Submitted to the Annals of Statistics

MINIMAX OPTIMAL VARIABLE CLUSTERING IN *G*-MODELS VIA CORD

By Florentina Bunea

Cornell University

By Christophe Giraud

Université Paris Sud

AND

By Xi Luo

Brown University

Abstract The goal of variable clustering is to partition a random vector $\mathbf{X} \in R^p$ in sub-groups of similar probabilistic behavior. Popular methods such as hierarchical clustering or K-means are algorithmic procedures applied to observations on $\mathbf{X}$, while no population level target is defined prior to estimation. We take a different view in this paper, where we propose and investigate model based variable clustering. We consider three models, of increasing level of complexity, termed generically G-models, with G standing for the partition to be estimated. Motivated by the potential lack of identifiability of the G-latent models, which are currently used in problems involving variable clustering, we introduce two new classes of models, the G-exchangeable and the G-block covariance models. We show that both classes are identifiable, for any distribution of $\mathbf{X}$.

Our focus is on clusters that are invariant with respect to unknown monotone transformations of the data, and that can be estimated in a computationally feasible manner. Both desiderata can be met if the clusters correspond to blocks in the copula correlation matrix of $\mathbf{X}$, assumed to have a Gaussian copula distribution. This motivates the introduction of a new similarity metric for cluster membership, CORD, and a homonymous method for cluster estimation. Central to our work is the derivation of the minimax value of the CORD cluster separation for exact partition recovery. We obtained the surprising result that this value is of order $\sqrt{\log(p)/n}$, irrespective of the number of clusters, or of the size of the smallest cluster. Our new procedure, CORD, available on CRAN, achieves this bound, is easy to implement and has computational complexity that is polynomial in p.

MSC 2010 subject classifications: Variable clustering, latent models, exchangeability, block covariance matrix, community estimation, minimax lower bound, consistent partition estimation,copula models, high dimensional models

1. Introduction. The problem of partitioning a random vector $\mathbf{X} =: (X_1, \ldots, X_p) \in \mathbb{R}^p$ in groups of variables that are similar in nature is known as variable clustering. Solutions to this problem are typically algorithmic in nature. Given $\mathbf{X}_1, \ldots, \mathbf{X}_n$ observations on $\mathbf{X}$, one defines a dissimilarity measures between the *data* vectors of observations on two components X_j and X_l of $\mathbf{X}$, and places the two variables in the same cluster if the *data* dissimilarity measure is small. The most popular measure is the Euclidean distance between vectors in $\mathbb{R}^n$ or, equivalently, the negative sample correlation and is typically the measure employed in popular clustering algorithms such as Hierarchical Clustering or K-means and their variants. We refer to [18] for an overview, and to clusters thus obtained as *estimated* clusters. Since no theoretical cluster target is postulated for $\mathbf{X}$, prior to estimation, it is unclear what the clusters obtained in this manner estimate.

In this work we take a different point of view, model-based clustering. We begin by considering models of probabilistic similarity between the components of $\mathbf{X}$ that will define *identifiable* target clusters, then use the data consisting in n observations on $\mathbf{X}$ to develop methods tailored to these targets. We perform a detailed theoretical study of the problem of exact cluster recovery. This provides a bridge for the existing gap in the literature on *variable* clustering.

We emphasize that our work is *not* on *data* clustering, for which a very large literature exists, centered around mixture models. We refer to [18] for a comprehensive discussion of related methods, and to [14] for foundational contributions. In data clustering, clusters are clouds of observations, and correspond to respective realizations of one of the mixture distributions, which is a distribution on the *whole* $\mathbb{R}^p$. In contrast, in variable clustering, the effort is to define cluster models relative to *subsets* of components X_j of $\mathbf{X} \in \mathbb{R}^p$, making it a completely different problem.

Our contributions and the organization of the paper are described below.

1: Identifiable models for variable clustering. In Section 2, we introduce and study three models that can be used for variable clustering. The parameter of interest, in each of these models, will be a partition $G = \{G_k\}_{1 \le k \le K}$ of $\{1, \ldots, p\}$, for some unknown integer K, and we will therefore refer to them, generically, as G-models. We introduce the following partial order on sets:

Partition partial (**PP**) order: Let $G = \{G_k\}_k$, $G' = \{G'_{k'}\}_{k'}$ be two partitions of $\{1, \ldots, p\}$. We say that G' is a sub-partition of G if for each k' there

exists k such that $G'_{k'} \subseteq G_k$. We define the partial order $\leq$ between two partitions G, G' of $\{1, \ldots, p\}$ by $G \leq G'$ if G' is a sub-partition of G.

We seek a minimal partition (according to the PP order) fulfilling some conditions of interest, called *model*. Since a partially ordered set can have multiple minimal elements, our model will be said to be *identifiable*, if there exists a unique minimal partition in the set of partitions following the model. Our first contribution is to define models for clustering that are identifiable in this sense, *without* any further distributional assumptions.

We first consider a model that is already popular for variable clustering, the G-latent model, which clusters components X_a of $\mathbf{X}$ that have the same latent generator Z_k. Although not analyzed theoretically in the manner proposed here, the G-latent models are widely used in applications ranging from biology to financial data. For instance, for analyses involving fMRI data, we refer to [24] and [8], and for a general overview to [25]. We show in Section 2.5.1 that one can construct vectors $\mathbf{X}$ that are latent with respect to two different minimal partitions. Thus, the G-latent models are not identifiable, in general, without further assumptions.

To address this issue, we introduce two new classes of models, of increasing complexity. The first is the class of G-exchangeable models, for which the within-group variables $(X_a)_{a \in G_k}$ are exchangeable within $\mathbf{X}$, but the between group variables are not. The second, larger, class of models that we propose is that of G-block covariance models, which are covariance models with a block-like structure: off-diagonal entries within a block are equal. We show in Sections 2.2 and 2.3, respectively, that both models are identifiable, with the latter being easier to estimate.

Moreover, in Section 2.5., in which we restrict ourselves to Gaussian distributions, we provide a full characterization of identifiability in Gaussian G-latent models. We also provide sufficient conditions under which the identifiable partitions for the three G-models under consideration coincide.

2: Identifiable clusters that are invariant with respect to monotone transformations of the data. Whereas the effort in Sections 2.2 and 2.3 is concentrated on defining models for clustering in full generality, care must be exercised when using them in real applications, as clusters must be invariant to simple scale or shift transformation of $\mathbf{X}$. Guided by the trade-off between distribution complexity and computational feasibility, we focus in this work on the class of what we term Gaussian copula G-models, introduced in Section 3. Within this class, we define, more generally, identifiable target clusters that are invariant relative to unknown monotone transforma-

tions of the data. With a view towards computational feasibility, our focus for the remainder of this article is on clusters that correspond to blocks of the copula correlation matrix R. To the best of our knowledge, our contributions *1* and *2* are new in the literature.

3: CORD*, a new metric for variable clustering.* Central to our work is a new, and simple, dissimilarity metric for variable clustering, invariant under monotone transformations of the data. Standard clustering techniques employ the correlation between two variables. In contrast, our metric is based on CORrelation Differences (CORD), a measure that attempts to capture how two variables behave relative to the rest, rather than relative to each other. We discuss this in detail in Section 4, where we also argue that MCORD, which is a measure of separation between clusters induced by CORD, can be used to define distinct clusters, even when the popular cluster separation measure, the within-between cluster correlation gap, is negative.

4: The minimax MCORD *lower bound for exact cluster recovery.* The strength of any variable clustering procedure is given by its ability to recover minimally separated clusters. Therefore, prior to developing an estimation procedure, it is crucial to assess the minimal size of the cluster separation below which no algorithm can be expected to recover consistently the correct partition. In Section 5 we present the main theoretical result of our work, the minimax MCORD separation rate for exact cluster recovery.

Minimax-type results are sometimes viewed as being overly pessimistic, since they are typically derived by considering very special distributions, corresponding to the worst case scenario for the problem under consideration. From this perspective our results are non-standard. Let $m =: min_{1\le k\le K}|G_k|$ denote the minimum cluster size. Theorem 2 shows that, for variable clustering, the minimax separation rate under the MCORD metric is $\sqrt{\log(p)/n}$, in a variety of cases, summarized below, *irrespective* of the size of K and m, or of the size of the correlation between variables in distinct groups:

- A case that is believed to be the most difficult for any clustering algorithm, corresponding to $K = p/2$ clusters of size $m = 2$ each. This is Proposition 8, in Appendix A.2.
- A case with in which we have $K = 2$, $m = 2$ and consequently the other cluster has size $p - 2$. This is Proposition 7, in Appendix A.2. This shows that having fewer clusters does not simplify the problem.
- A surprising case, with $K = 3$ and $m = p/3$, presented in Proposition 9 in Appendix A.2. This shows that the separation rate, under MCORD, is not influenced by m or K, as it is still $\sqrt{\log(p)/n}$, even though we have very

few, $K = 3$, clusters, each of large size, $p/3$.
• Moreover, our examples show that the same rate holds when variables in distinct groups are independent or when they are strongly correlated.

These results are, to the best of our knowledge, the first of their kind in the literature on clustering.

5: A new variable clustering algorithm, CORD. In Section 6 we develop CORD, a new clustering algorithm. It has moderate computational complexity, polynomial in p, and it is already freely available on-line, on CRAN. We prove that CORD can recover correctly the target partition, with high probability, at the MCORD minimax cluster separation rate. Section 7 contains summaries of our extensive numerical studies that show that CORD recovers clusters defined by our G-models, while existing variable clustering algorithms, such as K-means or hierarchical clustering based on the Euclidean distance, typically fail to do so. Section 7.4 provides further numerical evidence on the qualities of the newly introduced CORD metric as a bona-fide dissimilarity measure for variable clustering. We conclude with two real data examples, one of which is presented in the supplemental material.

1.1. *Contrast with clustering from network data.* The problem of grouping p entities in communities is also encountered in network analysis, and the estimation of these communities is known as the problem of community detection. However, the nature of the data is different. One observes a *network* of p objects, and associates with it a $p \times p$ adjacency matrix N: if there is a link between i and j, $N_{ij} = 1$, whereas the absence of a link yields $N_{ij} = 0$. The observed entries of N are assumed to be realizations of mutually independent Bernoulli random variables. One of the most basic models used to define communities in this context is the Stochastic Block Model (SBM), dating back to [15]. The model has been extensively studied in the past few years, and we give here only an incomplete list of references, most closely related to our work: [2], [21], [6], [11], [1], [20]. SBM assumes that $\mathbf{E}(N)$ has a block structure. In our work, we will employ the G-block correlation model, that postulates that the off-diagonal entries within a block of a correlation matrix R are equal. Since both problems involve block matrices, they may appear similar. However, there are fundamental differences between clustering based on SBM and clustering based on G-block correlation models, and we outline some of them below.

Data has different nature. The data generated from a SBM is a $p \times p$ binary matrix as above. The data generated from a G-model consists in n observations of a p dimensional vector, and can be organized as an $n \times p$ matrix of real numbers, with independent *rows*. We highlight two important repercussions of this difference.

Assume that in either SBM or the G-block correlation model the number of clusters K remains constant, as p increases. Assume also that n is fixed in the context of variable clustering, when p increases. Clearly, the clustering task in SBM becomes easier as p grows, since we have more data, whereas in G-models it deteriorates, as the quality of any estimator $\hat{R}$ of R deteriorates when p grows but n is set to a given value. Thus, while clustering a large network is statistically easier than clustering a small one, the opposite is true for variable clustering: clustering a high dimensional vector is harder than clustering a low dimensional vector, from the same amount n of observations.

Since N has independent entries and $\hat{R}$ has correlated (possibly strongly) entries, the statistical properties of the error terms involved in each problem, $V_1 =: \hat{R} - \mathbf{E}(\hat{R})$ and $V_2 =: N - \mathbf{E}(N)$, respectively, are very different. In SBM, cluster recovery typically requires control of the operator norm of V_2, see for instance [21]. We will show in Section 6 that, fortunately, for G-models one can construct consistent cluster estimators that only require control of the ℓ_∞ norm of V_1, which is of order $\sqrt{\log(p)/n}$, and not of the operator norm of V_1, which is of the order of $\sqrt{\|R\|_{op}\mathrm{trace}R}\sqrt{\log(pn)/n}$, see for instance [5]. Using the later may unnecessarily inflate the size of the minimal cluster separation needed for exact recovery.

Different cluster separation metric. Most SBM analyses focus on the case where there is a positive gap for the within-between community probabilities and investigate the minimal size of this gap for consistent clustering. As explained in the previous section, the quantity of interest for cluster recovery in G-models is not the correlation gap, but instead the newly introduced MCord separation. One of our main contributions is the investigation of the minimal MCord separation needed for successful clustering in G-models, which is our newly proposed metric for cluster separation.To the best of our knowledge, this question has not been posed in the SBM literature, and therefore not answered elsewhere.

2. Three models for variable clustering. Let $\mathbf{X} \in \mathbb{R}^p$ be a zero mean random vector. Let $G = \{G_k\}_{k=1,\dots,K}$ be a partition of $\{1, \dots, p\}$, for some integer K, $1 \leq K \leq p$. In what follows we introduce and compare three different models for the distribution of $\mathbf{X}$ that may potentially lead to dif-

ferent clusters G_k and different final partitions G. We begin by discussing partition identifiability.

2.1. *G-latent models.* Perhaps the most popular model connected to variable clustering, via the literature on clustering algorithms, is the simplest form of a latent variable model, as defined below. We will refer to this model as the G-latent model. Let G be a partition of $\{1,\dots,p\}$ into K groups and let $k : \{1,\dots,p\} \to \{1,\dots,K\}$ be an index assignment function defined by $G_k = \{a : k(a) = k\}$.

DEFINITION 1. **The G-latent model.** *The zero mean random vector* $\mathbf{X}$ *follows a G-latent model if there exists a zero mean random vector* $\mathbf{Z} = (Z_1,\dots,Z_K) \in R^K$ *such that* $X_a = Z_{k(a)} + E_a, \quad 1 \le a \le p$, *where the mutually independent zero mean error terms E_a have variance $\gamma_{k(a)}$, and are independent of the latent variables* $\mathbf{Z}$. *If a is a singleton, we use the convention $\gamma_{k(a)} = 0$.*

Let $\mathcal{L}(X) = \{G : \mathbf{X} \text{ is } G\text{-latent}\}$. The set $\mathcal{L}(X)$ is nonempty and finite, so it does have minimal elements G^ℓ. We will argue in Section 2.5.1 that, however, the minimal elements are not necessarily unique. Thus, without further assumptions, the G-latent models are not identifiable. This motivates the introduction and study of other classes of models for variable clustering, that are more flexible and are identifiable in full generality. We return to the class of G-latent models and offer sufficient conditions for their identifiability in Section 2.5.2.

2.2. *G-exchangeable models.* To introduce a new class of models for clustering, we will use the following intuition: two variables that belong to what will constitute a group in the partition should have similar behavior relative to all the other variables. One way to characterize this behavior is in terms of the joint distribution of $\mathbf{X} = (X_1,\dots,X_p)$. Below we will define partitions with the property that, in particular, two variables within the same group can be switched without changing the distribution of the vector, while switching variables between groups will change this distribution.

Formally, for a partition G, we define below the set S_G of permutations σ of $\{1,\dots,p\}$ that only permute elements within each group of the partition, but not between groups. Let $\mathcal{S}(G_k)$ be the set of permutations σ_k of G_k. Thus, $\mathcal{S}_G = \{\sigma = \sigma_1 \dots \sigma_K : \sigma_k \in \mathcal{S}(G_k)\}$. We let $\mathbf{X}_\sigma := (X_{\sigma(1)},\dots,X_{\sigma(p)})$.

DEFINITION 2. **The G-exchangeable model.** *The distribution of* $\mathbf{X}$ *is G-exchangeable (denoted by $\mathbf{X} \sim G$) if* $\mathbf{X}_\sigma \overset{law}{=} \mathbf{X}$, *for all $\sigma \in \mathcal{S}_G$.*

It is immediate to see that if $\mathbf{X}$ is G-latent, then $\mathbf{X}$ is G-exchangeable, but the reverse is not necessarily true, as the following example shows.

EXAMPLE 1. *Let* $\mathbf{X}$ *be a zero mean Gaussian vector with covariance matrix*

$$\Sigma = \begin{bmatrix} 1 & -1/2 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & -1/2 & 1 \end{bmatrix}$$

Then $\mathbf{X}$ *is* $G =: \{\{1,2\};\{2,3\}\}$*-exchangeable, but not* G*-latent. To see why notice that* Σ_{12} *and* Σ_{34} *are negative, and, for instance,* Σ_{12} *should have been positive, as it would be equal to* $var(Z_1)$ *if* $\mathbf{X}$ *were* G*-latent. Observe, moreover, that in this example the within-group correlations, which are all equal to* $-1/2$*, are smaller than between group correlations, which are all equal to zero.*

2.2.1. *G-exchangeable models are identifiable.* Let $\mathcal{E}(\mathbf{X}) = \{G : \ \mathbf{X} \sim G\}$. Observe that the set $\mathcal{E}(X)$ is not empty, as $\mathbf{X}$ is always G-exchangeable according to the *largest* possible partition with p groups, $G = \{\{1\}, \ldots, \{p\}\}$. Theorem 1 below shows that $\mathcal{E}(\mathbf{X})$ has a unique minimum, for a generic random vector $\mathbf{X}$, without any further distributional assumptions. We give a constructive proof of this result, that necessitates the definition of the minimum between two partitions. For any partition $G = \{G_k\}_{1 \le k \le K}$, we write $a \overset{G}{\sim} b$ if there exists $k \in \{1, \ldots, K\}$ such that $a, b \in G_k$. If for some a there does not exist $b \neq a$ such $a \overset{G}{\sim} b$, we call a a *singleton*.

DEFINITION 3. *For any two partitions* G, G' *of* $\{1, \ldots, p\}$*, we define* $G \wedge G'$ *as the partition induced by the equivalence relation* $a \overset{G \wedge G'}{\sim} b$ *iff there exist* $d \in \mathbb{N}$*, and* $c_1, \ldots, c_d \in \{1, \ldots, p\}$ *such that* $a \overset{G \text{ or } G'}{\sim} c_1 \overset{G \text{ or } G'}{\sim} \ldots \overset{G \text{ or } G'}{\sim} c_d \overset{G \text{ or } G'}{\sim} b$*. Informally,* $a \overset{G \wedge G'}{\sim} b$ *iff there exists a path through groups of* G *and* G' *which connects* a *to* b*.*

It is straightforward to check that $\overset{G \wedge G'}{\sim}$ is indeed reflexive, transitive and symmetric. With this definition, we have the following interesting properties that lead to the desired result, the identifiability of the newly introduced G-exchangeable models.

THEOREM 1. *Let* $\mathbf{X}$ *be a zero mean random vector with arbitrary distribution.*

1. *Consider $G \leq G'$. Then $X \sim G$ implies $X \sim G'$.*
2. *$G \wedge G' \leq G$ and $G \wedge G' \leq G'$*
3. *If $\mathbf{X} \sim G$ and $\mathbf{X} \sim G'$, then $\mathbf{X} \sim G \wedge G'$.*
4. *The set $\mathcal{E}(X)$ has a unique minimum, $G^\epsilon(X)$, with respect to the* **PP** *order.*

We refer to Appendix A.1 for a proof of this theorem. It shows that the minimum partition for which $\mathbf{X}$ is G-exchangeable always exists and its structure is intrinsic to the distribution of $\mathbf{X}$. The minimal partition $G^\epsilon(X)$ matches our intuition regarding what would constitute a reasonable clustering target. However, from a statistical point of view, estimating $G^\epsilon(X)$ may be challenging in high dimensions, as it will typically require estimation of high dimensional distributions and multiple testing. This motivates the introduction of a more general model, described in the following section.

2.3. *G-block covariance models.* Clustering according to G-exchangeable models may be too stringent, as it imposes conditions on the whole distribution of $\mathbf{X}$. In our third model, proposed below, we suggest clustering according to functionals of the distribution. In particular, we will consider the group structure on the covariance matrix induced by G-exchangeability.

DEFINITION 4. **G-block structure.** *Let G be a partition of $\{1, \ldots, p\}$. The covariance matrix Σ of $X \in \mathbb{R}^p$ is said to have a G-block structure if*

- *$var(X_a) =: \Sigma_{aa} = \Sigma_{bb} := var(X_b)$, if $a \overset{G}{\sim} b$*
- *$cov(X_a, X_{a'}) =: \Sigma_{aa'} = \Sigma_{bb'} := cov(X_b, X_{b'})$, for any $a \overset{G}{\sim} b$, $a' \overset{G}{\sim} b'$ and $a \neq a'$, $b \neq b'$.*

A $p \times p$ covariance matrix with K blocks has a reduced number of distinct parameters, $K(K-1)/2 + K$, as formalized in the result below. We write $|G_k|$ for the cardinality of G_k. With this notation, we list below a number of elementary, but very useful, properties of a covariance matrix with a G-block structure, proved in Section 1.1 of the supplemental material [4].

LEMMA 1. *A covariance matrix $\Sigma = cov(X)$ with a G-block structure has the following properties:*

1. *There exist $K(K-1)/2$ real numbers $c_{kk'} = c_{k'k}$, $1 \leq k, k' \leq K$, such that $\Sigma_{ab} = c_{kk'}$, for all $a \neq b$ with $k(a) = k$ and $k(b) = k'$.*
2. *There exist K non-negative numbers d_k such that, for all a, $\Sigma_{aa} = d_k$.*
3. *For all $1 \leq k \leq K$, we have $-\frac{d_k}{|G_k|-1} \leq c_{kk} \leq d_k$.*

2.3.1. *The G-block covariance models are identifiable.* For a vector $\mathbf{X} \in \mathbb{R}^p$ with covariance matrix Σ let $\mathcal{B}(X) = \{G : \ \Sigma \text{ has } G\text{-block structure}\}$. We show that G-block covariance models are identifiable by constructing the unique minimum element of $\mathcal{B}(X)$.

DEFINITION 5. *Let $G^{block}(X)$ be the partition induced by the equivalence relation: $a \overset{G^{block}}{\sim} b$ iff $var(X_a) = var(X_b)$ and $cov(X_a, X_c) = cov(X_b, X_c)$ for all $c \neq a, b$.*

It is immediate to verify that this relationship is reflexive, symmetric and transitive.

PROPOSITION 1. *The partition $G^{block}(X)$ is the minimum of $\mathcal{B}(X)$.*

We refer to Section 1.1 of the supplemental material [4] for a proof. Proposition 1 shows that, similarly to the G-exchangeable models, the class of the G-block covariance models is identifiable, for general distributions. Since $\mathcal{E}(X) \subseteq \mathcal{B}(X)$, we always have $G^{\epsilon}(X) \geq G^{block}(X)$, and the two partitions may not coincide, in general. However, there are situations where they do, and we discuss them in Section 2.5 below.

2.4. *Connections between the three G-models.* We have the following immediate relations between the three models introduced above. We omit the proof.

PROPOSITION 2. *Let $\mathbf{X}$ be a zero mean random vector in $\mathbb{R}^p$, with covariance matrix Σ. For any partition G of $\{1, \ldots, p\}$, we have:*
$\mathbf{X}$ is G-latent $\Rightarrow$ $\mathbf{X}$ is G-exchangeable $\Rightarrow$ Σ has G-block structure.

We show below that the three models are not equivalent. To see why, we first associate to a $p \times p$ G-block covariance matrix Σ a $K \times K$ matrix C defined below, using the notation introduced in Lemma 1.

DEFINITION 6. **The matrix of covariances C.** *Let Σ be a G-block covariance matrix. Let C be defined by*

- *$c_{kk'} = \Sigma_{ab}$, for all $a \neq b$ with $k(a) = k$ and $k(b) = k'$.*
- *if a is a singleton, with $k(a) = k$, we set $c_{kk} = \Sigma_{aa} = d_k$.*

The matrix $C = \{c_{kk'}\}_{1 \leq k, k' \leq K}$ is well-defined for any k, k' and it is symmetric. By Lemma 1 we always have $d_k \geq c_{kk}$, with equality for singletons. The matrix C collects the within and between group covariances and is not

always semi-positive definite. For instance, the matrix C associated with Example 1 is

$$C = \begin{bmatrix} -1/2 & 0 \\ 0 & -1/2 \end{bmatrix}$$

which is negative definite. There are however situations where C is positive semi-definite, as the following proposition shows. Its proof is immediate and therefore omitted.

PROPOSITION 3. *If $\mathbf{X}$ follows a G-latent model, then Σ has G-block structure with associated matrix $C = cov(\mathbf{Z})$. Therefore, in this case C is a semi-positive definite matrix and $\gamma_{k(a)} = d_{k(a)} - C_{k(a)k(a)} \geq 0$, for all a.*

REMARK 1. *The three G models are not equivalent: Proposition 3 shows that when Σ has a G-block structure, if the matrix C given by Definition 6 is not positive semi-definite, then $\mathbf{X}$ cannot admit a latent representation with the same G.*

2.5. *Gaussian G-models.* The theoretical results of the previous sections hold in full generality. More connections between the three G-models can be established if one makes distributional assumptions. In this section we study vectors $\mathbf{X}$ with a Gaussian distribution. In this case, we give a full characterization of identifiability in G-latent variable models, and connect this back with G-block covariance models and G-exchangeable models.

2.5.1. *The G-latent models may not be identifiable.* In this section we revisit the G-latent models and construct a counter-example that shows that the same vector $\mathbf{X}$ has two different latent representations, with respect to two different minimal partitions. In our construction, we will use the following simple lemma, proved in Section 1.1.3 of the supplement [4].

LEMMA 2. *If $\mathbf{X}$ has a mean zero Gaussian distribution with covariance matrix Σ that is G-block structured, and if the matrix C given by Definition 6 is positive semi-definite, then $\mathbf{X}$ follows a G-latent model.*

EXAMPLE 2. *Consider the positive definite matrix*

$$\Sigma = \begin{bmatrix} 1 & 0.25 & 0.26 & 0.26 \\ 0.25 & 1 & 0.26 & 0.26 \\ 0.26 & 0.26 & 1 & 0.25 \\ 0.26 & 0.26 & 0.25 & 1 \end{bmatrix}$$

and $\mathbf{X}$ *with Gaussian* $\mathcal{N}(0,\Sigma)$ *distribution. We have* $G^{block}(X) = \{\{1,2\};\{3,4\}\}$. *We use Definition 6 to construct the* C*-matrix associated to* Σ *relative to* $G^{block}(X)$

$$C(G^{block}(X)) = \begin{bmatrix} 0.25 & 0.26 \\ 0.26 & 0.25 \end{bmatrix}$$

which is not a positive semi-definite matrix. Therefore, by Proposition 3, $\mathbf{X}$ *cannot follow a* $G^{block}(X)$*-latent model. Consider now* $G = \{\{1\};\{2\};\{3,4\}\}$. *The* C*-matrix associated to* Σ *relative to* G *is*

$$C(G) = \begin{bmatrix} 1 & 0.25 & 0.26 \\ 0.25 & 1 & 0.26 \\ 0.26 & 0.26 & 0.25 \end{bmatrix}$$

which is positive semi-definite. Therefore, by (i) of Lemma 2, $\mathbf{X}$ *follows a* G*-latent model. By symmetry,* $\mathbf{X}$ *also follows a* G'*-latent model with* $G' = \{\{1,2\};\{3\};\{4\}\}$. *The two partitions* G *and* G' *are both minimal for the latent property with respect to the partial order* $\mathbf{PP}$ *and* $\mathcal{L}(X) = \{P : G \leq P \text{ or } G' \leq P\}$.

This example suggests that either $\mathbf{X}$ is $G^{block}(X)$-latent or its latent structure will contain singletons. This is indeed true, as formalized below.

PROPOSITION 4. *Assume* $\mathbf{X}$ *is a mean zero Gaussian vector. Let* G^ℓ *be a minimal element in* $\mathcal{L}(X)$. *Assume that* $\mathbf{X}$ *is not* $G^{block}(X)$*-latent. Then* G^ℓ *has at least one singleton.*

We refer to Appendix A.4 for a proof. A consequence of Proposition 4 is: If a minimal partition G^ℓ in $\mathcal{L}(X)$ has no singletons, then $G^\ell = G^{block}(X)$ and it is the unique minimum of $\mathcal{L}(X)$.

2.5.2. *Identifiable Gaussian G-models.* The following proposition summarizes the connection between Gaussian G-latent, G-exchangeable and G-block covariance models, and highlights a sufficient condition under which the minimum latent partition of $\mathbf{X}$ is $G^{block}(X)$, henceforth ensuring the uniqueness of the latent structure. We refer to Section 1.1.4 of the supplement [4] for a proof.

PROPOSITION 5. *Let* $\mathbf{X}$ *be a zero mean Gaussian distribution with covariance matrix* Σ. *Let* C *be the matrix associated with* Σ *and a partition* G *via Definition 6. Then: (i) For any partition* G *of* $\{1,\dots,p\}$ *we have that* $\mathbf{X}$ *is* G*-exchangeable* $\Leftrightarrow$ Σ *has a* G*-block structure. Thus,* $G^\epsilon(X) = G^{block}(X)$.

(ii) The following chain of equivalences holds: $\mathbf{X}$ *is* G*-exchangeable and* $C \geq 0 \Leftrightarrow \Sigma$ *has a* G*-block structure and* $C \geq 0 \quad \Leftrightarrow \mathbf{X}$ *is* G*-latent.*
(iii) If the matrix of covariances C *corresponding to the partition* $G^{block}(X)$ *is positive-semidefinite, then* $\mathbf{X}$ *is* $G^{block}(X)$*-latent, and this is the unique minimal partition according to which* $\mathbf{X}$ *admits a latent decomposition.*

3. Scale and translation invariant clustering via Gaussian copula *G*-models. If all the components of $\mathbf{X}$ have different variances, none of the proposed G-models would hold, despite possibly very strong reasons for grouping these components. Clearly, one should first make all components of $\mathbf{X}$ comparable at least in terms of scale and location, and postulate a model for clustering for a transformation of $\mathbf{X}$, not necessarily $\mathbf{X}$ itself.

Formally, we assume that there exists an unknown transformation h of $\mathbf{X}$ such that $\mathbf{Y} =: h(\mathbf{X})$ follows a G model. In Section 2 above we showed that the partition G is identifiable, under no further assumptions, if $\mathbf{Y}$ follows either a (a) G-exchangeable or a (b) G-block covariance model for clustering. In full generality, if we choose model (a), the estimation problem is difficult and possibly NP-hard, as it will involve estimation of high dimensional density functions and multiple testing in high dimensions. Moreover, not knowing the transformation h adds another layer of difficulty to an already complex problem.

We propose therefore to solve the problem for a particular class of distributions for $\mathbf{X}$, that of Gaussian copula distributions. We assume for the rest of this section and paper that $\mathbf{X}$ has a Gaussian copula distribution with zero mean, and copula function with parameters $\mu = 0$ and R, a correlation matrix, and we write $\mathbf{X} \stackrel{d}{=} \mathcal{C}(R)$. Recall that this implies that

$$\mathbf{Y} := (Y_1, \ldots, Y_p) := (h_1(X_1), \ldots, h_p(X_p)) =: h(\mathbf{X}) \stackrel{d}{=} \mathcal{N}_p(0, R), \tag{1}$$

with $h_a := \Phi^{-1} \circ F_a$, for each a, where Φ is the c.d.f of a standard Gaussian random variable, F_a is the marginal c.d.f. of X_a and $Var(Y_a) = 1$, for all a.

DEFINITION 7. *Assume that* $\mathbf{X}$ *has a Gaussian copula distribution with zero mean, and copula function with parameters* $\mu = 0$ *and* R*. If the Gaussian vector* $\mathbf{Y}$ *given by (1) follows a* G*-model, we say that* $\mathbf{X}$ *follows a Gaussian copula* G*-model.*

To emphasize that the covariance matrix of $\mathbf{Y}$ is a correlation matrix, whenever $\mathbf{Y}$ follow a G-block covariance model, we will call it *G-block correlation model.*

By Proposition 5 (i), the G-exchangeable and the G-block covariance models are equivalent for Gaussian vectors. Thus, either model assumption on $\mathbf{Y}$ will lead to *the same* identifiable target partition. Moreover, by Proposition 5 (iii), if the matrix C that can be associated with the correlation matrix R via Definition 6 is non-negative definite, G^{block} of R will also be the partition according to which $\mathbf{Y}$ has a latent decomposition.

Therefore, if $\mathbf{X}$ follows a Gaussian copula G-model and C is non-negative definite, we have the same target partition, irrespective of the particular type of G-model we postulate. Thus, if one prefers either the G-exchangeable or the G-latent model for their interpretability, one can use G-block correlation models for estimation, if it turns out that estimation is computationally efficient in that model. As we show in Section 6 below, this is indeed the case. We therefore focus for the rest of the paper on clustering via G-block correlation models, relative to the copula correlation matrix R.

4. The CORD similarity metric and its induced cluster separation metric. Classical clustering methods, such as K-means and hierarchical clustering, typically build clusters for which the within-group correlation is larger than that between groups. Therefore, if R has a block structure with associated matrix of correlations C, and c_{jk} denotes the correlation between variables in group G_j and those in G_k, one expects that consistent clustering would be possible with these classical clustering algorithms when the within-between correlation gap

$$\min_{j\neq k}(c_{jj} - c_{jk}) \tag{2}$$

is large enough. However, as our Example 1 shows, this requirement can be easily violated, and the gap (2) can be negative.

If clustering is done according to G-models, we argue below that the within-between correlation gap (2) is no longer the quantity that governs cluster separation. We define a new cluster separation metric that is positive, and can be large even if (2) is negative, thus offering an alternative to classical clustering.

To motivate our metric, notice that in G-block correlation models, if two variables with indices a and b are in the same group, then $R_{ac} = R_{bc}$, for any $c \neq a, b$. Conversely, if a and b are not in the same group, then $R_{ac} \neq R_{bc}$ for at least one variable $c \neq a, b$. Therefore, two variables X_a and X_b belong to the same cluster defined by the G^{block} partition of R if and only if

$$\max_{c\neq a,b} |R_{ac} - R_{bc}| = 0.$$

This motivates the introduction of a new disimilarity metric for variable clustering

$$\textsc{Cord}(a,b) := \max_{c\neq a,b} |R_{ac} - R_{bc}|,$$

which is tailored to G-models. The key quantity that quantifies the difficulty of clustering in G-models is the minimal CORD separation between two variables in distinct groups:

$$\textsc{MCord}(R) = \min_{a \overset{G^{block}}{\nsim} b} \textsc{Cord}(a,b). \tag{3}$$

Then, the larger the value of $\textsc{MCord}(R)$, the easier the cluster detection problem. We emphasize that requiring $\textsc{MCord}(R)$ be larger than η, for some $\eta > 0$, is in general much less stringent than requiring that the within-between correlation gap (2) be larger than η. To see why, notice that when the partition has no singletons, we always have

$$\textsc{MCord}(R) \geq \min_{j\neq k}(c_{jj} - c_{jk}).$$

Referring again to Example 1, it may happen that the right-hand side is negative, while the left-hand side is large.

In Section 5 below we study the minimax optimal separation size for consistent partition estimation with respect to the newly introduced $\textsc{MCord}(R)$. The investigation of the minimax optimal correlation gap (2) ensuring consistent partition estimation, is considered in [3].

5. The minimax lower bound for cluster separation with respect to MCord(R). We always have $\textsc{MCord}(R) > \eta$, with $\eta = 0$. However, a larger value of η will be needed for retrieving consistently the partition G^{block} from noisy observations. We give its minimax optimal value below. To begin with, we define a general class of models over which we establish our minimax lower bounds. Let R be the copula correlation matrix of $\mathbf{X}$, assumed to have an arbitrary G^{block} structure. We define

$$\mathcal{R}(\eta) \quad = \quad \{R : \ \textsc{MCord}(R) > \eta\}.$$

Our goal is to find the minimal value η^* below which exact recovery with high-probability is impossible. Formally, we will show that

$$\inf_{\widehat{G}} \sup_{R\in\mathcal{R}(\eta)} \mathbf{P}_R(\widehat{G} \neq G^{block}) \geq \text{constant} > 0,$$

for all $0 \leq \eta < \eta^*$, where the infimum is taken over all possible estimators.

THEOREM 2.

$$\textit{For any } \ 0 \le \eta < \eta^* := \frac{0.6\sqrt{\frac{\log(p)}{n}}}{1+0.6\sqrt{\frac{\log(p)}{n}}},$$

we have

$$\inf_{\widehat{G}} \sup_{R\in\mathcal{R}(\eta)} \mathbf{P}_R(\widehat{G} \neq G^{block}) \ge \frac{1}{2e+1} \ge \frac{1}{7},$$

where the infimum is taken over all possible estimators.

We observe that $\eta^* \sim 0.6\sqrt{\frac{\log(p)}{n}}$ when $\log(p) = o(n)$, as announced in the discussion in the Introduction. The proof is given in the Appendix A.2 and consists in the collection of the proofs of Propositions 7, 8 and 9, each showcasing a different scenario for the number of clusters, the size of the smallest cluster and the correlation between and within clusters. In all these cases, the minimax cluster separation rate is $\sqrt{\log(p)/n}$, a new and surprising result in the literature on variable clustering.

6. Consistent estimation of minimally separated clusters via Cord. In this section with present an algorithm that achieves the separation lower bound, for Gaussian copula G-block correlation models. We recall that we place variables X_a and X_b in the same group when their transforms Y_a and Y_b are in the same group, and the latter happens if and only if $\textsc{Cord}(a,b) := \max_{c\neq a,b} |R_{ac} - R_{bc}| = 0$, where R is the copula correlation matrix of $\mathbf{X}$. Since CORD does not use the size of the correlation of a pair of variables to decide on group membership, estimation of clusters in G-block models cannot be performed in the standard manner of thresholding an estimator of R, followed by taking the connected components of the resulting graph. Instead, our estimator will be based on thresholding an estimator of CORD, in an iterative fashion, as we explain below. To estimate CORD, we only require estimates of the entries of R from n observed independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$. Let $\widehat{R}$ be an estimator of R. We first recall that, under the Gaussian copula model on $\mathbf{X}$,

$$R_{ab} = Cov(Y_a, Y_b) = Corr(Y_a, Y_b) = \sin(\frac{\pi}{2}\tau_{ab}),$$

where Kendall's tau coefficient is defined by

$$\tau_{ab} = \mathbf{E}\left[\text{sign}((Y_a - \tilde{Y}_a)(Y_b - \tilde{Y}_b))\right] = \mathbf{E}\left[\text{sign}((X_a - \tilde{X}_a)(X_b - \tilde{X}_b))\right],$$

with $\tilde{X}_a$ ($\tilde{Y}_a$) and $\tilde{X}_b$ ($\tilde{Y}_b$) independent copies of X_a (Y_a) and X_b (Y_b), respectively. Therefore, we can estimate R_{ab} by

$$\widehat{R}_{ab} = \sin(\frac{\pi}{2}\widehat{\tau}_{ab}), \tag{4}$$

where

$$\widehat{\tau}_{ab} = \frac{2}{n(n-1)} \sum_{1\le i<i'\le n} \text{sign}(X_a^i - X_a^{i'})\text{sign}(X_b^i - X_b^{i'}).$$

If we have reasons to believe that $\mathbf{X}$ has a Gaussian distribution, then we can alternatively consider as an estimator of R the sample correlation matrix. Given the entries of $\widehat{R}$, the goal is to estimate G^{block}, the minimal partition with respect to which R has a block structure. To a given estimator $\widehat{R}$ of R we associate the estimator

$$\widehat{\text{CORD}}(a,b) := \max_{c\neq a,b} |\widehat{R}_{ac} - \widehat{R}_{bc}|, \quad a,b \in \{1,\dots,p\}\,, \tag{5}$$

of the CORD metric. We then estimate the partition $\widehat{G}$ according to the following algorithm.

Algorithm 1

- Input: $\widehat{R}$ and $\alpha > 0$
- Initialization: $S = \{1,\dots,p\}$ and $\widehat{\text{CORD}}(a,b) = \max_{c\neq a,b} |\widehat{R}_{ac} - \widehat{R}_{bc}|$ and $l = 0$
- Repeat: while $S \neq \emptyset$
  - $l \leftarrow l + 1$
  - If $|S| = 1$ Then $\widehat{G}_l = S$
  - If $|S| > 1$ Then
    - $(a_l, b_l) = \underset{a,b\in S,\ a\neq b}{\text{argmin}}\ \widehat{\text{CORD}}(a,b)$
    - If $\widehat{\text{CORD}}(a_l,b_l) > \alpha$ Then $\widehat{G}_l = \{a_l\}$
    - If $\widehat{\text{CORD}}(a_l,b_l) \le \alpha$ Then $\widehat{G}_l = \left\{ s \in S : \widehat{\text{CORD}}(a_l,s) \wedge \widehat{\text{CORD}}(b_l,s) \le \alpha \right\}$
  - $S \leftarrow S \setminus \widehat{G}_l$
- Output: the partition $\widehat{G} = (\widehat{G}_l)_{l=1,\dots,k}$

We emphasize that this algorithm does not require as input the specification of the number K of groups. The algorithmic complexity for computing $\widehat{R}$ defined by (4) is $O(p^2 n \log(n))$ (see [7]) and the complexity of Algorithm 1 is $O(p^3)$, so the overall complexity of our estimation procedure is $O\big(p^2(p \vee n\log(n)\big)$. In the following, we provide conditions ensuring that $\widehat{G} = G^{block}$.

PROPOSITION 6. *Let* $\mathbf{X}$ *be a zero mean random vector with a Gaussian copula distribution with parameter* R. *Let* $\widehat{R}$ *be any estimator of* R. *We define* $\tau = |\widehat{R} - R|_\infty$ *and we consider two parameters* (α, η) *fulfilling*

$$\alpha \geq 2\tau \quad \textit{and} \quad \eta \geq 2\tau + \alpha. \tag{6}$$

Then, if $R \in \mathcal{R}(\eta)$, *our algorithm yields* $\widehat{G} = G^{block}$.

We refer to Appendix A.3 for a proof of this proposition and to Section 1.2 of the supplemental material [4] for the proof of the following corollary.

COROLLARY 1. *Let us consider* (α, η) *fulfilling*

$$\alpha \geq 4\pi\sqrt{\frac{(1+A)\log(p)}{n}} \quad \textit{and} \quad \eta \geq 4\pi\sqrt{\frac{(1+A)\log(p)}{n}} + \alpha,$$

for some $A > 0$. *If* $\mathbf{X}$ *has a zero mean Gaussian copula distribution and* $R \in \mathcal{R}(\eta)$, *then the output of our algorithm applied to the estimator* $\widehat{R}$ *given by (4), is consistent:* $\widehat{G} = G^{block}$, *with probability higher than* $1 - p^{-2A}$.

REMARK 1. Therefore, with $\tau \leq$ constant $\times \sqrt{\frac{\log(p)}{n}}$, thresholding $\widehat{\text{CORD}}$ at level 2τ guarantees exact recovery, whenever the CORD separation η is at least 4τ. Whereas the exact values of the constants will depend on the particular parameters of the distribution, Theorem 2 in the previous section guarantees that the order of the CORD separation is tight, since we showed that there exists constant* such that if the CORD separation equals $\eta^* =$ constant$^* \times \sqrt{\frac{\log(p)}{n}}$, exact recovery is impossible, by any method.

If $\mathbf{X}$ has a Gaussian distribution, we can use the sample correlation matrix for $\widehat{R}$. The results of Corollary 1 above remain unchanged, modulo constants, and are presented in Corollary 1 of Section 1.3 of the supplement [4].

7. Simulations.

7.1. *Simulation design.* In this section we verify numerically our theoretical findings. We consider a number of types of G-block covariance matrices, of increasing level of complexity. To construct them, we recall Lemma 1 of Section 3.1, that shows that the G-block structure for Σ is specified by $C = (c_{kk'})$ and $D = (d_k)$, $1 \leq k, k' \leq K$. We note that once Σ is constructed for our simulation models, the correlation matrix R has the same G-block structure.

We consider the following models for Σ corresponding to the following matrix C:

- Model 1: C is block diagonal and each block is a 2×2 block matrix B, where $B_{11} = 0.6$, $B_{22} = 2$, $B_{12} = B_{21} = 0.8$.
- Model 2: $C = B^T B$ where B is a random $K \times K$ matrix with independent entries. Each entry takes the value $+1$ and -1 with equal probability $0.5 \times K^{-1/2}$, and the value 0 with probability $1 - K^{1/2}$.
- Model 3: $C = \left(B - \bar{B}\right)^T \left(B - \bar{B}\right)$ where B is randomly generated as in Model 2 and $\bar{B}_{kk'} = K^{-1} \sum_{k=1}^{K} B_{kk'}$.

The matrix C is positive definite in Models 1 and 2, and it is positive semidefinite in Model 3. We have also considered negative definite matrices C by subtracting a small positive number from all the diagonal entries of the C in Model 3, and the results are similar. In all these models, we set $d_k = c_{kk} + 1$, $k = 1, \ldots, K$, to ensure that the resulting Σ (and R) is positive definite. We set $K = 10$ to be the the number of equal-size groups of variables, for each of two representative cases $p = 200$ and $p = 1600$. For each p, we run simulations for samples of sizes $n = 100, 200, 300, \ldots, 1000$. All simulations are repeated 100 times.

We simulate the data from either a Gaussian or Gaussian copula distribution. In the Gaussian setting, the observations on $\mathbf{X}$ are simulated from a $\mathcal{N}(0, R)$ distribution, with R having the block structure given by Models 1 - 3. To simulate Gaussian copula observations, we apply transformation $f(x|v_k)$ to each variable X_a if $a \in G_k$, for $\mathbf{X}$ simulated from $\mathcal{N}(0, R)$. The transformation functions take the form

$$f(x|v_k) = \log\left(1 + |x - v_k|^{1/3}\right) \cdot \operatorname{sign}(x - v_k)$$

where the parameters v_k, $k = 1, \ldots, K$, are iid from a uniform distribution between $(-1, +1)$. The distribution of each variable after transformation is bimodal. We obtained very similar results in the Gaussian case and the Gaussian copula case. For brevity, we only report the result for the Gaussian

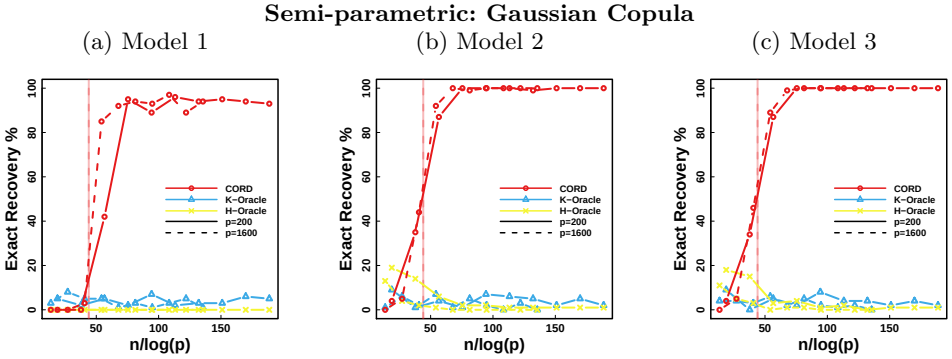
copula case in this section and we refer to Section 2 of the supplemental material [4] for the results in the Gaussian case.

7.2. *G-Block Estimation.* The goal of our algorithm is to create sub-groups of vectors of dimension n, from a given collection of p vectors of observations, each of dimension n, according to CORD. This task can be viewed as that of clustering p objects in $\mathbb{R}^n$. The existing clustering algorithms are not tailored to recovering groups with this structure, but they can serve as comparative methods. We thus compare the performance of CORD with two popular clustering algorithms: K-means and Hierarchical Clustering (HC). We apply K-means on the columns of the $n \times p$ matrix of scaled observations, and we use the negative correlation as distance matrix in HC. Both Kmeans and HC require specification of the number of groups K. We use the true $K = 10$ in these methods to evaluate their oracle performance. In our simulations, both CORD and HC use estimates of Pearson's correlation under the Gaussian setting, and of the sinusoid transform of Kendall's tau in (4) under the Gaussian copula setting. In the following simulation study, we use the fixed threshold $\alpha = 2\sqrt{\log(p)/n}$, since our theoretical results suggest the usage of a threshold proportional to $\sqrt{\log(p)/n}$ with a constant larger than 1. We also provide a data-dependent choice in the next section. We consider the set-up described above, with $n = 100, 200, \ldots, 1000$ and $p = 200, 1600$.

7.2.1. *Exact recovery.* We first compare the performance CORD and its competitors in terms of exact true group recovery. Figure 1 shows the percentages of exact recovery by K-means, HC, and CORD. CORD clearly outperforms both K-means and HC when n is about 400 or larger in all the models. K-means and HC, even with the oracle choice of K and large $n = 1000$, fail to recover the true groups exactly. It is also interesting to note that their recovery percentages are flat around 0 as n increases, and the percentages for HC decrease for small n in Model 2 and 3.

We proved in Section 6 that, whenever the group separation parameter η and the threshold level α are both larger than appropriate multiples of $n^{-1/2}\log^{1/2} p$, Algorithm 1 recovers consistently the block structure of R. Figure 1 shows that CORD recovers the true groups almost 100% of the runs when the condition $\eta \geq \alpha + 2\tau = 2n^{-1/2}\log^{1/2} p + 2\tau$ is satisfied. This condition translates into saying that the sample size is larger than a minimal sample size, shown as a vertical line in our plots. The recovery percentages for CORD show sharp rise around the calculated lower bounds, and the curves for $p = 200$ and $p = 1600$ almost overlap. This confirms empirically the phase transition by the scaling rate $n^{-1/2}\log^{1/2} p$, observed in our converge rates in Corollary 1 and Theorem 2.

Figure 1: Percentages of exact recovery by K-means (K-Oracle, medium blue lines, triangle points), HC (H-Oracle, light yellow lines, cross points) and CORD (dark red lines, circle points) across 100 runs, when $p = 200$ (solid lines) and $p = 1600$ (dashed lines). The minimal sample sizes are shown by light red vertical lines. All standard errors are smaller than 5%.

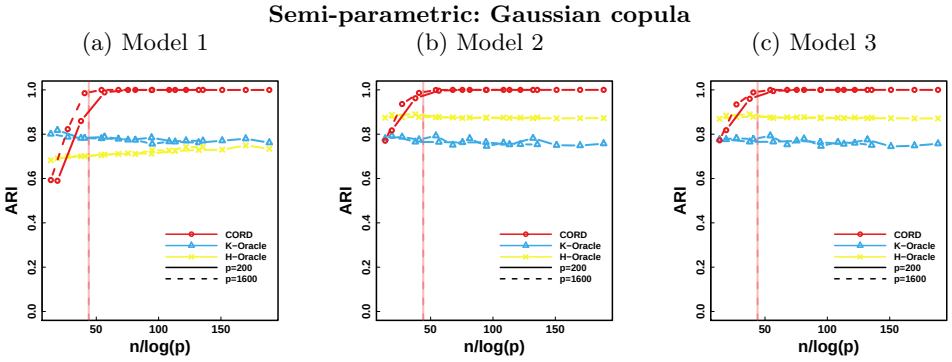


7.2.2. *Approximate recovery.* Perfect recovery is a strong requirement for any partition estimation algorithm, so we also investigate the performance of CORD compared to K-means and HC in terms of partial recovery. We measure the partial recovery performance in terms of the Adjusted Rand Index (ARI) [17], which is a continuous metric with values in $[-1, +1]$ for comparing two partitions of a set. Two identical partitions yield an ARI value of 1, and random partitions are expected to yield a zero value. Figure 2 shows that, in terms of partial recovery, our method is again a strong competitor of existing clustering algorithms in any regime: the ARI of CORD, a method that also estimates the number of communities, grows steeply in the small n regime, and becomes 1 fast, whereas the other algorithms, which require the number of groups as an input, level off at an ARI of approximately 0.8, even when the sample size increases substantially.

7.3. *Performance of* CORD *with data-driven* α.

7.3.1. *Cross validation for the* CORD *threshold.* The threshold α appearing in Corollary 1 does not involve any unknown quantity and they can be used directly. In practice, however, it is advisable to use a data-driven choice of the threshold for CORD. We propose to use the following type of cross-validation for this purpose. The idea is to construct a loss function over a grid of the α values, such that the value of α for which this loss has minimum value is also the value of α for which we have consistent recovery of our communities.

Figure 2: Average ARI by K-means (K-Oracle, medium blue lines, triangle points), HC (H-Oracle, light yellow lines, cross points) and CORD (dark red lines, circle points) across 100 runs.



To this end, it is useful to introduce the following operator on correlation matrices.

Given a correlation matrix R and a partition G, we introduce a block averaging operator $\Upsilon(R, G)$ which produces a G-block structured matrix of the same size as R. For any $a \in G_k$ and $b \in G_{k'}$, the output matrix entry $[\Upsilon(R, G)]_{ab}$ is given by

$$[\Upsilon(R,G)]_{ab} = \begin{cases} |G_k|^{-1}\left(|G_k|-1\right)^{-1}\sum_{i,j\in G_k, i\neq j} R_{ij} & \text{if } a \neq b \text{ and } k = k' \\ |G_k|^{-1}\,|G_{k'}|^{-1}\sum_{i\in G_k, j\in G_{k'}} R_{ij} & \text{if } a \neq b \text{ and } k \neq k' \\ 1 & \text{if } a = b. \end{cases}$$

In essence, this operator averages over the submatrix of R with indices in G_k and $G_{k'}$ respectively, except for the diagonal entries. It produces an estimator of R that has the G block structure.

Given a matrix loss function L, and an estimator of R, our cross-validation (CV) procedure is as follows:

For a sample of size n on $\mathbf{X}$, we first randomly split it into two equal-size datasets: the training dataset and the validating dataset. We then compute the corresponding correlation estimates $\hat{R}_t$ and $\hat{R}_v$, with indices referring to the training and validating datasets, respectively.

For given $D > 1$, and each value α_l, $l = 1, \ldots, D$, on a grid, we use Algorithm 1 to compute $\hat{G}_l$, using $\hat{R}_t$ and α_l as input. Our data dependent threshold is the grid value $\hat{\alpha}$ given by

$$\hat{\alpha} = \underset{\alpha_l}{\operatorname{argmin}}\, L\left(\hat{R}_v, \Upsilon\left(\hat{R}_t, \hat{G}_l\right)\right), \tag{7}$$

TABLE 1
Average α values selected by our cross validation approach and the clustering performance of the cross validated α measured by the exact recovery percentages (ERP) and average ARI, under either the Gaussian (G) or Gaussian copula (GC) setting. All standard errors are smaller than 0.06.

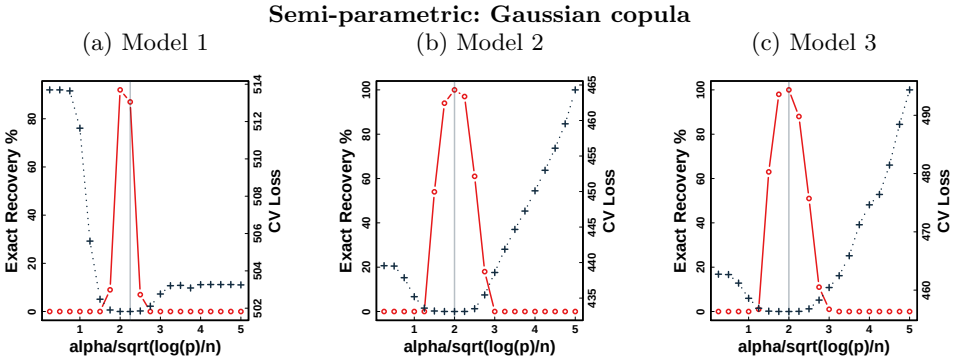
	Model 1		Model 2		Model 3	
	G	GC	G	GC	G	GC
$\alpha/\left(n^{-1/2}\log^{1/2} p\right)$	2.025	2.208	1.540	1.630	1.545	1.595
ERP	88	81	100	100	100	100
ARI	0.980	0.964	1.000	1.000	1.000	1.000

In our numerical studies, we use the Frobenius loss $L(R, M) = \|R - M\|_F$. In the Gaussian setting, $\widehat{R}$ is the sample correlation matrix, in the Gaussian copula setting it is sinusoid transformed Kendall's tau in (4).
The theoretical investigation of this criterion is beyond the scope of this work, and we present below its performance on a simulated example.

7.3.2. *Numerical studies.* We use the case $n = 1000$ and $p = 1600$ to study the "large n, large p" performance of our cross validation approach. The training dataset and validating dataset have the same sample size 500 respectively. Since the theoretical value of α is proportional to $n^{-1/2}\log^{1/2} p$, we use a grid of $\alpha/\left(n^{-1/2}\log^{1/2} p\right) = 0.25, 0.5, \ldots, 5$. Table 1 shows that the average α selected by CV is between $1.5\times n^{-1/2}\log^{1/2} p$ and $2.2\times n^{-1/2}\log^{1/2} p$, close to our fixed choice $2 \times n^{-1/2}\log^{1/2} p$ in the previous simulation study. Using the cross validated α in each run, CORD recovers exactly the true groups over 80% of the runs for Model 1, and 100% for Model 2 and 3. Under Model 1, CORD still achieves high average ARI values, larger than 0.964. The performance metrics under Model 1 decrease only slightly for the Gaussian Copula setting compared with the Gaussian setting.
We show the relationship between the average CV losses and exact recovery percentages in Figure 3. This shows that the optimal ranges of α values for high exact recovery percentages are also associated with low average CV losses. All these ranges are around $2n^{-1/2}\log^{1/2} p$ in all models.

7.4. *The* CORD *metric for correct community placement of pairs of variables.* The CORD method can be viewed as a method for clustering the p rows (and columns) of $\widehat{R}$. Most clustering methods are applied relative to a given distance or metric, where the metric is chosen such that pairs of objects are grouped together if it is small, and placed in different groups if the metric is large.
In this section we validate numerically that the CORD *metric* given by (5) can

Figure 3: Average CV losses (black dotted lines, plus points) and exact recovery percentages (red solid lines, circle points) across 100 runs. The minimal average CV losses are shown by gray vertical lines.

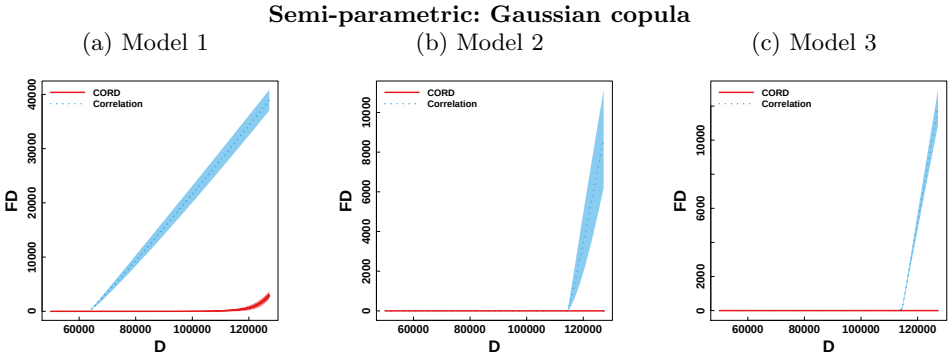


indeed be viewed as a bona fide variable clustering metric, when communities have nontrivial size, larger or equal to 2, and are generated from the G-models introduced and studied in the previous sections. The study presented below is therefore not of the CORD algorithm per se, but rather of its metric, which is the crucial ingredient of our method. To assess correct community assignment of pairs of variable, we compare the number of false discoveries to that of discoveries (we define below what each means in this context). Given the data, we pair two variables (a, b) if $\widehat{\text{CORD}}(a, b) < v$, where v is a given threshold value. We call such a pair a "discovery". We recall that $\widehat{\text{CORD}}$ is given by (5), with $\widehat{R}$ an appropriate estimator of R. Of these discoveries, some are false, in that X_a and X_b do not belong to the same group in the true partition. Formally, we let

$$FD(v) = |\{\widehat{\text{CORD}}(a, b) < v,\ a \overset{G^\epsilon(X)}{\nsim} b,\ a > b\}|, \tag{8}$$
$$D(v) = |\{\widehat{\text{CORD}}(a, b) < v,\ a > b\}|.$$

Both $FD(v)$ and $D(v)$ increase with v, and we want to study the effect of increasing v on the number of false discoveries FD relative to the number discoveries D. Figure 4 compares the average number of false discoveries FD as we vary the threshold v such that the percentage of discoveries D reaches up to 10% of the total number of entries in the lower triangular part of the correlation matrix. The results correspond to $n = 500$ and $p = 1600$, chosen as a representative case. This figure shows that the CORD metric

Figure 4: Average numbers of false discoveries (FD) and discoveries (D) after thresholding either the CORD (red solid lines) or correlation (blue dotted lines) metrics. The shaded areas show mean±SD.



yields almost zero false positives, for a relatively high range of the threshold values.

We also compared FD and D with $\widehat{\text{CORD}}$ in (8) replaced by the estimated negative correlation, which is either the Pearson's correlation in the Gaussian setting, or the negative Kendall's tau in the Gaussian copula setting. We recall that the former is equivalent with using the ℓ_2 distance between observations on two variables. These metrics are the ingredients of the variable clustering algorithms used for comparison in the previous two subsections. Figure 4 shows that for our G-latent models, the negative correlation (Pearson's or Kendall's tau) metric can yield as many as 40 000 false positives out of a total of over 120 000 discoveries. This suggests strongly that the commonly used metrics, irrespective of the algorithm that employs them, are not optimal for G-exchangeable, G-block and G-latent models, whereas the CORD metric adapts to this structure.

8. The analysis of Standard & Poor 100 Stocks. To illustrate the applicability of the G-exchangeable models for defining communities of variables, we apply CORD for the analysis of a stock dataset, in order to study which stocks are grouped together. We compare the resulting partitions with those obtained from the K-means and HC algorithms, respectively. This dataset contains daily returns of the stocks listed in the Standard & Poor 100 index (as of March 21, 2014), from January 1, 2006 to December 31, 2008. The stocks with incomplete data, due to company reorganization for example, are removed from the analysis, and this leaves 91 stocks with

returns from 755 trading days in total.
We first compare the groups recovered by CORD, K-means, and HC. We use the cross validation approach of Section 7.3.1 to select α, and it selects the threshold $\alpha = 1.2n^{-1/2}\log^{1/2} p$ with $K = 30$. We thus set $K = 30$ in Kmeans and HC (using the negative Kendall's tau distance) to compare the clustering performance. Table 2 lists the stocks in the first 5 CORD groups. Because the partitions from K-means and HC are different, we list all the groups from K-means and HC that contain these stocks for each CORD group. It is clear that CORD groups together directly competing companies that offer similar or almost identical services, while K-means and HC can group companies that provide different kinds of services and products. For example, Home Depot and Lowe's are the only two companies in a CORD group, probably because they are both home improvement stores. However, Starbucks is added to this group by both K-means and HC, even though Starbucks is a coffee shop chain. Further comparison is provided in Section 3 of the supplemental material [4].

An additional application of the CORD method to an fMRI data analysis is presented in Section 4 of the supplemental material [4].

Acknowledgement. The research of F. Bunea was supported in part by NSF-DMS 1310119; the research of C. Giraud was supported in part by Labex LMH ANR-11-IDEX-003-02; and the research by X. Luo was supported in part by NIH P01AA019072, P20GM103645, P30AI042853, R01NS052470, and S10OD016366.

REFERENCES

[1] Arias-Castro, E. and Verzelen, N. (2013) Community detection in random networks. Preprint.

[2] Amini, A.A., Chen, A., Bickel, P.J., and Levina, E. (2013) Pseudo likelihood metgods for community detection in large sparse networks. Annals of Statistics 41 (4), 2097 - 2122.

[3] Bunea, F., Giraud, C., Royer, M. and Verzelen, N. (2016) Near minimax optimal variable clustering via convex optimization. Preprint.

[4] Bunea, F., Giraud, C., and Luo, X, (2015). Supplement to: Minimax optimal variable clustering in G-models via CORD, Preprint.

[5] Bunea, F. and Xiao, L, (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. Bernoulli 21(2), 1200-1230.

[6] Chen, Y. and Xu, J. (2015) Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices, JMLR, to appear.

[7] Christensen, D. (2005). Fast algorithms for the calculation of Kendall's τ. Computational Statistics 20 (1): 51–62.

TABLE 2

The first 5 groups of stocks found by CORD *and the groups from K-means and HC that contain these stocks. K-means has two groups (separated by a horizontal line) that contain the stocks in Group 1 of* CORD*, and HC merges Group 4 and 5 of* CORD *into one group along with other stocks.*

Group	Industry Section	CORD	K-means	HC
1	Oil & Gas Equipment and Service	Anadarko Petroleum, Devon Energy, Halliburton, National Oillwell Varco, Occidental Petroleum, Schlumberger	Anadarko Petroleum, Devon Energy, Halliburton, National Oillwell Varco, Occidental Petroleum, Schlumberger, Apache, ConocoPhillips, Chevron, Exxon --- National Oillwell Varco	Anadarko Petroleum, Devon Energy, Halliburton, National Oillwell Varco, Occidental Petroleum, Schlumberger, Apache, ConocoPhillips, Chevron, Exxon, Freeport-McMoran
2	Telecom	AT&T, Verizon	AT&T, Verizon, Pfizer, Merck, Lilly, Bristol-Myers	AT&T, Verizon
3	Railroads	Norfolk Southern, Union Pacific	Norfolk Southern, Union Pacific	Norfolk Southern, Union Pacific, Du Pont, Dow Chemical, Monsanto
4	Home Improvement	Home Depot, Lowe's	Home Depot, Lowe's, Starbucks	Home Depot, Lowe's, Starbucks, Costco, Target, Wal-Mart, FedEx, United Parcel Service
5	Air Freight & Logistics	FedEx, United Parcel Service	FedEx, United Parcel Service, Caterpillar, Du Pont, Dow Chemical, Emerson Electric, General Electric, Honeywell, 3M, Nike, United Technologies	Home Depot, Lowe's, Starbucks, Costco, Target, Wal-Mart, FedEx, United Parcel Service

[8] Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human brain mapping, 33(8), 1914-1928.

[9] Craddock, R. C., Jbabdi, S., Yan, C. G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., ... and Milham, M. P. (2013). Imaging human connectomes at the macroscale. Nature methods, 10(6), 524-539.

[10] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping, 2(4), 189-210.

[11] Gao,C., Ma, Z., Zhang, A. y. and Zhou, H. (2015) Achieving optimal misclassification proportion in stochastic block model, Preprint.

[12] Giraud, C. (2014) Introduction to High-Dimensional Statistics. Chapman and Hall.

[13] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. JASA **58**: 13–30.

[14] Hartigan, J.A. (1975) Clustering algorithms. John Wiley & Sons.

[15] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983) Stochastic blockmodels: First steps. Social Networks 5, 109- 137.

[16] Hornik, K. (2005). A clue for cluster ensembles. Journal of Statistical Software, 14(12).

[17] Hubert, L., and Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.

[18] Izenman, A. (2008) Modern Multivariate Statistical Techniques. Springer Text in Statistics.

[19] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[20] Klopp, O., Tsybakov, A. B. and Verzelen, N. (2015) Oracle inequalities for network models and sparse graphon estimation.

[21] Lei, J. and Rinaldo, A. (2015) Consistency of spectral clustering in stochastic block models. Annals of Statistics 43(1), 215-237.

[22] Massart, P. (2007) Concentration Inequalities and Model Selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics 1896, Springer.

[23] Murtagh, F., and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.

[24] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., ... and Petersen, S. E. (2011). Functional network organization of the human brain. Neuron, 72(4), 665-678.

[25] Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006). Learning the structure of linear latent variable models. The Journal of Machine Learning Research, 7, 191-246.

[26] Simmonds, D. J., Pekar, J. J., and Mostofsky, S. H. (2008). Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. Neuropsychologia, 46(1), 224-232.

[27] Xue, G.; Aron, A. R. and Poldrack, R. A.(2008). Common neural substrates for inhibition of spoken and manual responses Cerebral Cortex, Oxford Univ Press,, 18, 1923-1932.

APPENDIX A: PROOFS

A.1. Identifiability in G-exchangeable models: Proof of Theorem 1 . The first two statements follow by the definition of G-exchangeability and that of $G \wedge G'$.

We begin by proving the third claim. For this, we first show that any permutation $\sigma \in \mathcal{S}_{G\wedge G'}$ can be decomposed into transpositions $\sigma = \tau_1 ... \tau_d$ with each transposition τ_i permuting two elements a_i, b_i belonging to the same group of G or G'. Below, we write τ_{ab} for the transposition between a and b. We need the following lemma.

LEMMA 3. *Let $c_1, \dots, c_k$ be $k \geq 2$ distinct points. Then the transposition $\tau_{c_1 c_k}$ can be decomposed into $\tau_{c_1 c_k} = \prod_{i=1}^{N} \tau_{c_{j_i} c_{j_i+1}}$ with $N \leq 2k$ and $1 \leq j_i \leq k-1$.*

Proof of the lemma. For $k = 2$ it is obvious. For $k = 3$ we have $\tau_{c_1 c_3} = \tau_{c_2 c_3} \tau_{c_1 c_2} \tau_{c_2 c_3}$. Assume that the property holds up to k elements. We have $\tau_{c_1 c_{k+1}} = \tau_{c_k c_{k+1}} \tau_{c_1 c_k} \tau_{c_k c_{k+1}}$ and applying the induction hypothesis to $\tau_{c_1 c_k}$ we obtain the property at level $k+1$. The proof of the lemma is complete. □

We first observe that a permutation $\sigma \in \mathcal{S}_{G\wedge G'}$ can be decomposed into transpositions $\sigma = \tau_1 ... \tau_\ell$, each transposition acting on a single class $(G \wedge G')_k$. For any $a, b \in (G \wedge G')_k$ we have $a \overset{G \text{ or } G'}{\sim} c_1 \overset{G \text{ or } G'}{\sim} \dots \overset{G \text{ or } G'}{\sim} c_d \overset{G \text{ or } G'}{\sim} b$ so according to the lemma, the transposition τ_{ab} can be decomposed as a product of transpositions, where each transposition is between two elements which are in a same G_k or G'_k. So, any permutation $\sigma \in \mathcal{S}_{G\wedge G'}$ can be decomposed into transpositions permuting elements belonging to a same group of G or G'.
Let us conclude. If τ is a transposition between two elements of a common cluster G_k or G'_k, then $\mathbf{X}$ and $\mathbf{X}' = \mathbf{X}_\tau$ have the same distribution by hypothesis. So, for any permutation σ, the variables $\mathbf{X}_\sigma$ and $\mathbf{X}'_\sigma = \mathbf{X}_{\tau\sigma}$ have the same distribution. So, by induction, using the above decomposition, we obtain that for any permutation $\sigma \in \mathcal{S}_{G\wedge G'}$, the variables $\mathbf{X}_\sigma$ and $\mathbf{X}$ share the same distribution. Hence $X \sim G \wedge G'$. This concludes the proof of the third claim.

We conclude the proof of this proposition by proving the fourth claim. The set $\mathcal{E}(X)$ is non-empty since the trivial partition $G = \{\{1\}, \dots, \{p\}\}$ belongs to $\mathcal{E}(X)$. It is also a finite set, and we can enumerate it $\mathcal{E}(X) = \{G_1, \dots, G_M\}$. Define the sequence $G'_1, \dots, G'_M$ recursively according to

- $G'_1 = G_1$,
- $G'_k = G_k \wedge G'_{k-1}$ for $k = 2, \ldots, M$.

According to Proposition 1, we have by induction that $G'_1, \ldots, G'_M \in \mathcal{E}(X)$. In addition, we have both $G'_k \leq G'_{k-1}$ and $G'_k \leq G_k$, so by induction $G'_k \leq G_1, \ldots, G_k$. Hence, the partition $G^{\epsilon}(X) := G'_M = G_1 \wedge G_2 \wedge \ldots \wedge G_{M-1}$ is the minimum of $\mathcal{E}(X)$. □

A.2. Minimax lower bounds: Proof of Theorem 2. Our proofs are based on a version of Fano's lemma suited for our application. Specifically, we will employ Birgé's Lemma (Corollary 2.18 in [22]), which we state below, translated to our problem.

LEMMA 4. *For any partition estimator $\widehat{G}$, and for any collection of distinct $G^{(j)}$-block covariance matrices $R^{(j)} \in \mathcal{R}(\eta)$, $1 \leq j \leq M$, we have*

$$\begin{aligned}
\sup_{R \in \mathcal{R}(\eta)} \mathbf{P}_R(\widehat{G} \neq G^{block}) &\geq \max_{j=1,\ldots,M} \mathbf{P}_{R^{(j)}}(\widehat{G} \neq G^{(j)}) \\
&\geq \frac{1}{2e+1} \wedge \left(1 - \max_{j \geq 2} \frac{\mathcal{K}(R^{(j)}, R^{(1)})}{\log(M)}\right), \qquad (9) \\
&\geq \frac{1}{2e+1} \wedge \left(1 - \max_{j \geq 2} \frac{n\|(R^{(1)})^{-1}(R^{(j)} - R^{(1)})\|^2}{2\log(M)}\right), \qquad (10)
\end{aligned}$$

where $\mathcal{K}(R^{(j)}, R^{(1)})$ denotes the Kulback-Leibler divergence between two Gaussian likelihoods based on n observations.

PROOF. Inequality (9) is Birgé's Lemma (Corollary 2.18 in [22]), translated to our problem. We prove inequality (10) below. We only need to check that

$$\mathcal{K}(R^{(j)}, R^{(1)}) \leq n\|(R^{(1)})^{-1}(R^{(j)} - R^{(1)})\|^2/2. \qquad (11)$$

We have

$$\begin{aligned}
\mathcal{K}(R^{(j)}, R^{(1)}) &= \frac{n}{2}\left(Trace((R^{(1)})^{-1}R^{(j)} - I) - \log det\left((R^{(1)})^{-1}R^{(j)}\right)\right) \\
&= \frac{n}{2}(F((R^{(1)})^{-1}R^{(j)}) - F(I)).
\end{aligned}$$

with $F(S) = Trace(S) - \log det(S)$. Notice that F is convex, and therefore

$$F(I + H) - F(I) \leq \langle I - (I + H)^{-1}, H \rangle,$$

since the gradient of our function F in $I+H$ is $I-(I+H)^{-1}$. Let $\sigma_1 \geq \sigma_2 \geq \ldots$ be the singular values of H. Then

$$\langle I-(I+H)^{-1}, H\rangle = \sum_k \frac{\sigma_k^2}{1+\sigma_k} \leq \sum_k \sigma_k^2 = \|H\|^2.$$

Consequently, (10) holds, and the proof of this lemma is complete. □

The task therefore is that of constructing M correlation matrices in $\mathcal{R}(\eta)$ and determining η^* for which either (9) or (10) in Lemma 4 is at least $\frac{1}{2e+1}$. Whereas the proof of Theorem 2 would be completed by constructing one instance of these M matrices, we will provide three such constructions, in order to emphasize the generality of our results. We begin by constructing a sequence of matrices with $K=2$ blocks, one of size 2 and one of size $p-2$.

PROPOSITION 7. *Let us fix $0 \leq \alpha < \beta < 1$. We construct a sequence of correlation matrices in $\mathcal{R}(\eta)$ as follows. To any pair $a \neq b$ in $\{1,\ldots,p\}$, we associate the matrix $R^{(a,b)}$ with 1 on the diagonal and α off the diagonal, except for positions (a,b) and (b,a) where we define $R^{(a,b)}_{ab} = R^{(a,b)}_{ba} = \beta$. The partition $G^{block(a,b)}$ associated to $R^{(a,b)}$ is then $\{\{a,b\};\{1,\ldots,p\}\setminus\{a,b\}\}$. When*

$$\frac{\beta-\alpha}{1-\beta} \leq \sqrt{\frac{\log(p(p-1)/2)}{(2+e^{-1})n}}, \tag{12}$$

then

$$\inf_{\widehat{G}} \sup_{R\in\mathcal{R}(\eta)} \mathbf{P}_R(\widehat{G} \neq G^{block}) \geq \inf_{\widehat{G}} \max_{a\neq b} \mathbf{P}_{R^{(a,b)}}(\widehat{G} \neq G^{block(a,b)}) \geq \frac{1}{2e+1}\,.$$

PROOF OF PROPOSITION 7. First, notice that the cardinality of the set $\{R^{(a,b)} : a \neq b \in \{1,\ldots,p\}\}$ is $M = p(p-1)/2$, and all matrices $R^{(a,b)} \in \mathcal{R}(\eta)$, for any $\eta \leq \beta-\alpha$. We denote by J the matrix with all entries equal to 1, which is positive semi-definite. The matrix $R^{(1,2)}-\alpha J$ has $1-\alpha$ on the diagonal, 0 off the diagonal, except in $(1,2)$ and $(2,1)$ where it takes value $\beta-\alpha$. The lowest eigenvalue of $R^{(1,2)}-\alpha J$ is $1-\beta$, hence $|(R^{(1,2)})^{-1}|_{op} \leq (1-\beta)^{-1}$. We also observe that $\|R^{(1,2)}-R^{(a,b)}\|_F^2 = 4(\beta-\alpha)^2$ for any $(a,b) \neq (1,2)$. Hence

$$\|(R^{(1,2)})^{-1}(R^{(1,2)}-R^{(a,b)})\|_F^2 \leq 4\left(\frac{\beta-\alpha}{1-\beta}\right)^2.$$

Then, when (12) holds, inequality (10) of Lemma 4 above gives the desired result:

$$\inf_{\widehat{G}} \max_{a\neq b} \mathbf{P}_{R^{(a,b)}}(\widehat{G} \neq G^{block(a,b)}) \geq \frac{1}{2e+1} \wedge \left(1 - \frac{2n(\beta-\alpha)^2}{(1-\beta)^2 \log(p(p-1)/2)}\right) \geq \frac{1}{2e+1}.$$

□

We now construct a second sequence of matrices with many small blocks. In this example, we have $K = p/2$ blocks of size 2, assuming that p is even.

PROPOSITION 8. *Let $R^{(1)}$ be the matrix $R^{(1)} = \alpha J + diag(A, \dots, A)$ where J is the matrix with all entries equal to 1 and*

$$A = \begin{pmatrix} 1-\alpha & \beta-\alpha \\ \beta-\alpha & 1-\alpha \end{pmatrix}.$$

The partition $G^{block(1)}$ associated to $R^{(1)}$ is $G^{block(1)} = \{\{1,2\}; \dots; \{p-1,p\}\}$. We denote by τ_{ab} the transposition between a and b in $\{1,\dots,p\}$ and by $R^{(a,b)}$ the matrix $R^{(a,b)} = \left[R^{(1)}_{\tau_{ab}(i)\tau_{ab}(j)}\right]_{i,j=1,\dots,p}$ built from $R^{(1)}$ by permuting the rows and columns indexed by a and b. The partition $G^{block(a,b)}$ associated to this matrix is the same as $G^{block(1)}$ except that the indices a and b have been switched. Then, when

$$\frac{\beta-\alpha}{1-\beta} \leq \sqrt{\frac{\log(p(p-2)/4)}{(4+2e^{-1})n}}, \tag{13}$$

no estimator can perfectly recover the partition G^{block} uniformly on the set of distributions $\{\mathbf{P}_{R^{(1)}}\} \cup \{\mathbf{P}_{R^{(a,b)}} : a = 2k,\ k = 1,\dots,p/2,\ b > a\}$, with probability larger than $2e/(2e+1)$.

PROOF OF PROPOSITION 8. In this case, the cardinality of the set

$$\left\{R^{(a,b)} : a = 2k,\ k = 1,\dots,p/2,\ b > a\right\} \cup \left\{R^{(1)}\right\}$$

is $M = p(p-2)/4+1$, and $R^{(a,b)} \in \mathcal{R}(\eta)$ for any α and β such that $\eta \leq \beta-\alpha$. As in the previous example, since the lowest eigenvalue of A is $1-\beta$, we have

$|(R^{(1)})^{-1}|_{op} \leq (1-\beta)^{-1}$. We also observe that $\|R^{(1)} - R^{(a,b)}\|_F^2 = 8(\beta-\alpha)^2$. Therefore, when (13) holds, Lemma 4 gives

$$\inf_{\widehat{G}} \max_{a\in\{2,4,\ldots,p\},\ b>a} \left(\mathbf{P}_{R^{(a,b)}}(\widehat{G} \neq G^{block(a,b)}) \vee \mathbf{P}_{R^{(1)}}(\widehat{G} \neq G^{block(1)})\right)$$
$$\geq \frac{1}{2e+1} \wedge \left(1 - \frac{4n(\beta-\alpha)^2}{(1-\beta)^2 \log(p(p-2)/4)}\right) \geq \frac{1}{2e+1}.$$

□

By analogy with clustering via SBM, we may expect that when all clusters have a common size m, the clustering problem become easier and easier as m grows. The following example shows that it is not the case in general, even when m grows linearly with p.

PROPOSITION 9. *Let $0 < \epsilon < 1$ and define*

$$C(\epsilon) = \begin{bmatrix} \epsilon & \epsilon - \epsilon^2 & -\epsilon \\ \epsilon - \epsilon^2 & \epsilon & \epsilon \\ -\epsilon & \epsilon & 2 \end{bmatrix},$$

which is positive semi-definite. Consider the following covariance matrix with $K = 3$ blocks of equal size $m = p/3$ given by $\Sigma := AC(\epsilon)A^T + I$, where A is a $p \times 3$ hard assignment matrix given by $A_{j1} = 1$, if $1 \leq j \leq m$, and zero otherwise, $A_{j2} = 1$, if $m+1 \leq j \leq 2m$, and zero otherwise and $A_{j3} = 1$, if $2m+1 \leq j \leq 3m$, and zero otherwise. For $a \in \{1,\ldots,m\}$ and $b \in \{m+1,\ldots,2m+1\}$, we construct $\Sigma^{(a,b)}$ by permuting the rows and columns indexed by a and b. We collect the $M = m^2+1$ candidate matrices in a set $\mathcal{M}$. Let G'^{block} be the partition associated with one of the generic matrices $\Sigma' \in \mathcal{M}$. If

$$\epsilon \leq \sqrt{\frac{0.8\log(p/3)}{n}} \quad \text{and} \quad \eta \leq 0.8\epsilon,$$

then

$$\inf_{\widehat{G}} \sup_{R\in\mathcal{R}(\eta)} \mathbf{P}_R(\widehat{G} \neq G^{block}) \geq \inf_{\widehat{G}} \max_{\Sigma'\in\mathcal{M}} \mathbf{P}_{\Sigma'}(\widehat{G} \neq G'^{block})$$
$$\geq \frac{1}{2e+1}.$$

PROOF OF PROPOSITION 9. We begin by noting that Σ given above is not a correlation matrix, but since $Diagonal(\Sigma) = (1+\epsilon, ..., 1+\epsilon, 3, ..., 3)$, its associated correlation matrix R will have the same block structure. Moreover,

since $\mathrm{MCord}(\Sigma) = 2\epsilon$, then $\mathrm{MCord}(R) = c\epsilon$, for some $c \in [2/\sqrt{6}, 2/\sqrt{3}]$, and $R \in \mathcal{R}(\eta)$, for any $\eta \leq 0.8\epsilon$. To avoid the notational clutter that would be introduced by normalizing Σ, we will work in what follows with covariance matrices, which does not affect the proof.

We will once again appeal to Lemma 4, and we calculate below the KL-divergence $\mathcal{K}(\Sigma^{(a,b)}, \Sigma)$. By symmetry, we can assume that $a = 1$ and $b = m+1$. We write henceforth Q for the permutation matrix associated to the transposition of 1 and $m+1$ and $\Sigma' = Q\Sigma Q^T$. We start by writing out the eigenvalues and the corresponding eigenvectors of $C(\epsilon)$:

- $\lambda_1 = \epsilon(2-\epsilon)$ with $u_1^T = \frac{1}{\sqrt{2}}[1 \quad 1 \quad 0]$
- $\lambda_2 = 2 + \epsilon^2$ with $u_2^T = \frac{\epsilon}{2\sqrt{1+\epsilon^2/2}}[-1 \quad 1 \quad 2/\epsilon]$
- $\lambda_3 = 0$ with $u_3^T = \frac{1}{\sqrt{2+\epsilon^2}}[1 \quad -1 \quad \epsilon]$

Let $v_k^T = \frac{1}{\sqrt{m}}(Au_k)^T$, for $k = 1, 2$. Then, with I denoting the $p \times p$ identity matrix, we have

$$\Sigma = \sum_{k=1}^{2} m\lambda_k v_k v_k^T + I, \qquad \Sigma^{-1} = -\sum_{k=1}^{2} \frac{m\lambda_k}{1+m\lambda_k} v_k v_k^T + I$$
$$\Sigma' - \Sigma = \sum_{k=1}^{2} m\lambda_k [Qv_k v_k^T Q^T - v_k v_k^T].$$

We have $Qv_k = v_k + \Delta_k$ with $\Delta_k^T = \frac{1}{\sqrt{m}}[(u_k)_2 - (u_k)_1, 0...0, (u_k)_1 - (u_k)_2, 0...0]$. Then, $\Delta_1 = 0$, $\Delta_2^T = \frac{\epsilon}{\sqrt{1+\epsilon^2/2}\sqrt{m}}[1, 0...0, -1, 0...0]$, and

$$\Sigma' - \Sigma = m\lambda_2(v_2\Delta_2^T + \Delta_2 v_2^T + \Delta_2\Delta_2^T).$$

We notice that $v_1 v_1^T \Delta_2 = 0$ and $v_1 v_1^T v_2 = 0$, so putting pieces together:

$$\Sigma^{-1}\Sigma' = I + \Sigma^{-1}(\Sigma' - \Sigma) = I + m\lambda_2 M,$$

where

$$M := (I - \frac{m\lambda_2}{1+m\lambda_2} v_2 v_2^T)(v_2\Delta_2^T + \Delta_2 v_2^T + \Delta_2\Delta_2^T).$$

Writing $s := \Delta_2^T v_2$, $\gamma := \Delta_2^T\Delta_2$ and $\rho := m\lambda_2/(1+m\lambda_2)$, we have

$$\begin{aligned} M &= [v_2\Delta_2^T + \Delta_2 v_2^T + \Delta_2\Delta_2^T - \rho(v_2\Delta_2^T + sv_2 v_2^T + sv_2\Delta_2^T)] \\ &= v_2\Delta_2^T(1 - \rho(1+s)) + \Delta_2 v_2^T + \Delta_2\Delta_2^T - \rho s v_2 v_2^T. \end{aligned}$$

Let us compute the eigenvalues of M. We observe that the range of M is spanned by v_2 and Δ_2, so we seek for eigenvectors $\omega = v_2 + \alpha\Delta_2$. Since

$$s = -\frac{2\epsilon^2}{m(2+\epsilon^2)} = -\frac{2\epsilon^2}{m\lambda_2}, \quad \text{and} \quad \gamma = \frac{4\epsilon^2}{m\lambda_2} = -2s, \tag{14}$$

computing $M\omega$, we obtain

$$\begin{aligned} M\omega &= [(1-\rho(1+s))(s+\alpha\gamma) - \rho s(1+\alpha s)]v_2 + [1+\alpha s + s + \alpha\gamma]\Delta_2 \\ &= [(1-\rho(1+s))(1-2\alpha) - \rho(1+\alpha s)]sv_2 + [1+(1-\alpha)s]\Delta_2, \end{aligned}$$

so

$$\begin{aligned} M\omega = \mu\omega &\iff \begin{cases} 1+(1-\alpha)s = \alpha\mu \\ [(1-\rho(1+s))(1-2\alpha) - \rho(1+\alpha s)]s = \mu \end{cases} \\ &\iff \begin{cases} \alpha = \frac{1+s}{\mu+s} \\ 0 = \mu^2 + \mu[\rho s(2+s)] + (1-\rho)(s+2)s. \end{cases} \end{aligned}$$

Therefore, the two non-zero eigenvalues μ_1 and μ_2 of M fulfill

$$\begin{cases} \mu_1 + \mu_2 = -\rho s(2+s) \\ \mu_1\mu_2 = (1-\rho)s(2+s). \end{cases}$$

Since $\Sigma^{-1}\Sigma' = I + m\lambda_2 M$, with M of rank 2, we can now compute $\mathcal{K}(\Sigma', \Sigma)$:

$$\begin{aligned} \frac{2}{n}\mathcal{K}(\Sigma', \Sigma) &= \operatorname{Tr}(m\lambda_2 M) - \log\det(I + m\lambda_2 M) \\ &= m\lambda_2(\mu_1+\mu_2) - \log(1 + m\lambda_2(\mu_1+\mu_2) + (m\lambda_2)^2\mu_1\mu_2) \\ &= -m\lambda_2\rho s(2+s) - \log\left(1 - m\lambda_2\rho s(2+s) + (m\lambda_2)^2(1-\rho)s(2+s)\right). \end{aligned}$$

We observe that $m\lambda_2(1-\rho) = \rho$, so the terms in the log cancel and finally, plugging the values in (14) into the above display, we obtain

$$\mathcal{K}(\Sigma', \Sigma) = -\frac{n}{2}m\lambda_2\rho s(2+s) = 2n\rho\epsilon^2\left(1 - \frac{\epsilon^2}{m(2+\epsilon^2)}\right) \leq 2n\epsilon^2.$$

Since we have $M \geq (p/3)^2$ candidate matrices, Lemma 4 ensures that when $\epsilon \leq \sqrt{\frac{2e\log(p/3)}{(2e+1)n}}$ the probability of mis-clustering is larger than $1/(2e+1)$.

□

A.3. Proof of consistent recovery. *Proof of Proposition 6.* First, we notice that, since $|\widehat{R} - R|_\infty = \tau$, we have

$$|R_{ac} - R_{bc}| - 2\tau \le |\widehat{R}_{ac} - \widehat{R}_{bc}| \le |R_{ac} - R_{bc}| + 2\tau,$$

for any a, b, c. Hence $\widehat{\textsc{Cord}}(a,b) - 2\tau \le \textsc{Cord}(a,b) \le \widehat{\textsc{Cord}}(a,b) + 2\tau$. We then observe that $a \overset{G^{block}}{\sim} b \implies \textsc{Cord}(a,b) = 0 \implies \widehat{\textsc{Cord}}(a,b) \le 2\tau$, and when the group separation condition $R \in \mathcal{R}(\eta)$ holds

$$a \overset{G^{block}}{\nsim} b \implies \textsc{Cord}(a,b) > \eta \implies \widehat{\textsc{Cord}}(a,b) > \eta - 2\tau.$$

In particular, under the condition (6) and the group separation condition $R \in \mathcal{R}(\eta)$, we have

$$a \overset{G^{block}}{\sim} b \iff \widehat{\textsc{Cord}}(a,b) \le \alpha. \tag{15}$$

Let us prove the proposition by induction on l. We consider the algorithm at some step l and assume that the algorithm was consistent up to this step, i.e. $\widehat{G}_j = G^{block}_{k(a_j)}$ for $j = 1, \dots, l-1$.

If $|S| = 1$, then it directly follows that $\widehat{G} = G^{block}$. Assume now that $|S| > 1$.

(i) If $\widehat{\textsc{Cord}}(a_l, b_l) > \alpha$, then according to (15) no $b \in S$ is in the same group as a_l. Since the algorithm has been consistent up to step l, it means that a_l is a singleton and $\widehat{G}_l := \{a_l\} = G^{block}_{k(a_l)}$.

(ii) If $\widehat{\textsc{Cord}}(a_l, b_l) \le \alpha$, then $a_l \overset{G^{block}}{\sim} b_l$ according to (15). The equivalence (15) furthermore ensures that $\widehat{G}_l = S \cap G^{block}_{k(a_l)}$. Since the algorithm has been consistent up to this step we have $G^{block}_{k(a_l)} \subset S$ and hence $\widehat{G}_l = G^{block}_{k(a_l)}$. To conclude, the algorithm remains consistent at step l and the proposition follows by induction. □

A.4. Proofs for *G*-latent models.

A.4.1. *Proof of Proposition 4.* Let G^ℓ be minimal in $\mathcal{L}(X)$. Since $\mathcal{L}(X) \subseteq \mathcal{B}(X)$, then G^ℓ is a sub-partition of $G^\beta := G^{block}(X)$, that is $G^\beta \le G^\ell$. Write k^β, respectively k^ℓ for the index function associated to the partition G^β, respectively G^ℓ. Similarly, we write C^β for the C-matrix associated to G^β according to Definition 6 and C^ℓ for the one associated to G^ℓ. According to the very definition of C^β and C^ℓ, for every $a \ne b$, we have

$$C^\beta_{k^\beta(a)k^\beta(b)} = \Sigma_{ab} = C^\ell_{k^\ell(a)k^\ell(b)}. \tag{16}$$

Since $\mathbf{X}$ is not G^β-latent, the matrix C^β is not positive semi-definite: There exists $x^\beta \neq 0$ such that $(x^\beta)^T C^\beta x^\beta < 0$. We next show that it enforces G^ℓ to have at least one singleton.
Let $K^\beta \times K^\beta$ be the size of C^β and $K^\ell \times K^\ell$ be the size of C^ℓ. Since G^ℓ is a sub-partition of G^β, up to a relabelling of the groups in G^ℓ, we have

$$G_1^\beta = G_1^\ell \cup \ldots \cup G_{K_1}^\ell \ , \ldots , G_{K^\beta}^\beta = G_{K_1+\ldots+K_{K^\beta-1}+1}^\ell \cup \ldots \cup G_{K_1+\ldots+K_{K^\beta}}^\ell .$$

We can associate to the vector $x^\beta \in R^{K^\beta}$ the vector $x^\ell \in R^{K^\ell}$ defined by

$$x^\ell = (x_1^\beta, 0, \ldots, 0, x_2^\beta, 0, \ldots, 0, \ldots)$$

where x_k^β is located at index $i(k) = K_1 + \ldots + K_{k-1} + 1$. For any $k \neq j$, taking $a \in G_{i(k)}^\ell$ and $b \in G_{i(j)}^\ell$, Equation (16) gives $C_{kj}^\beta = C_{i(k)i(j)}^\ell$. Assume now that G^ℓ has no singleton: For any $k \in \{1, \ldots, K^\beta\}$ we can choose $a \neq b$ in $G_{i(k)}^\ell$. For this choice of a and b, the Equation (16) gives $C_{kk}^\beta = C_{i(k)i(k)}^\ell$. So $C_{kj}^\beta = C_{i(k)i(j)}^\ell$ for all $k, j \in \{1, \ldots, K^\beta\}$. Hence, we obtain $(x^\ell)^T C^\ell (x^\ell) = (x^\beta)^T C^\beta x^\beta < 0$. This is impossible since $\mathbf{X}$ is G^ℓ-latent, so C^ℓ is the covariance matrix of the latent variables and hence it is positive (semi-)definite. We conclude that G^ℓ has at least one singleton. □

Department of Statistical Science
Cornell University
Ithaca, NY 14853-2601, USA
E-mail: fb238@cornell.edu

Département de Mathématiques
Université Paris-Sud
F-91405 Orsay Cedex, France
E-mail: christophe.giraud@math.u-psud.fr

Department of Biostatistics
and Center for Statistical Science
Brown University
Providence, RI 02912, USA
E-mail: xi.rossi.luo@gmail.com

Submitted to the Annals of Statistics

SUPPLEMENT TO: MINIMAX OPTIMAL VARIABLE CLUSTERING IN *G*-MODELS VIA CORD

By Florentina Bunea

Cornell University

By Christophe Giraud
Université Paris Sud
and

By Xi Luo
Brown University

Abstract The following document provides proofs of some of the lemmas, propositions and corollaries in [3]. It also contains additional simulation results and an additional data analysis. The numbering of lemmas, propositions and corollaries will be the same as in [3]. Results in [3] will be used without explicit reference.

1. Supplemental proofs and theoretical results.

1.1. *Proofs for Section 2.*

1.1.1. *Proof of Lemma 1 of Section 2.3.* The first two properties follow directly from Definition 4. The third statement follows by a simple calculation, since Σ_{G_k}, the covariance matrix of $(X_a)_{a\in G_k}$, has d_k on the diagonal and c_{kk} outside of the diagonal, and hence $d_k - c_{kk}$ and $d_k + (|G_k|-1)c_{kk}$ as eigenvalues. Since Σ_{G_k} is positive semi-definite, its eigenvalues are non-negative, which gives the last inequalities. □

1.1.2. *Proof of Proposition 1 of Section 2.3.* 1) We first show that $G^\beta := G^{block}(X)$ belongs to $\mathcal{B}(X)$. By Definition 5, we have that $var(X_a) = var(X_b)$ for $a \overset{G^\beta}{\sim} b$. Also, for $a \overset{G^\beta}{\sim} b$, $a' \overset{G^\beta}{\sim} b'$ with $a \neq b$, $a' \neq b'$, and $a \neq a'$, $b \neq b'$, we have

$$cov(X_a, X_{a'}) = cov(X_b, X_{a'}) = cov(X_b, X_{b'})$$

if $b \neq a'$. If $b = a'$ and $a \neq b'$ then $a \sim a' = b \sim b'$ in G^β so $cov(X_a, X_{a'}) = cov(X_a, X_b) = cov(X_a, X_{b'})$ since $b \overset{G^\beta}{\sim} b'$ and $a \neq b, b'$. If $b = a'$ and $a = b'$ the equality is obvious. Therefore, Definition 4 is met and we conclude that Σ is G^β-block structured.

2) Next, we show that $G^\beta \leq G$, for any $G \in \mathcal{B}(X)$. For this, we need to show that whenever $a \overset{G}{\sim} b$, then $a \overset{G^\beta}{\sim} b$, which will establish that G is a sub-partition of G^β. If $a \overset{G}{\sim} b$ we have, by Definition 4:

- $var(X_a) = var(X_b)$, since $a \overset{G}{\sim} b$.
- $cov(X_a, X_c) = cov(X_b, X_c)$, for any $c \neq a, b$, since $a \overset{G}{\sim} b$ and $c \overset{G}{\sim} c$.

Therefore, $a \overset{G^\beta}{\sim} b$, which concludes the proof. □

1.1.3. *Proof of Lemma 2 of Section 2.5.1.* Let $\mathbf{Z}$ be a Gaussian $\mathcal{N}(0, C)$ random variable and E_a be independent random variables, independent of $\mathbf{Z}$ and with $\mathcal{N}(0, \Sigma_{aa} - C_{k(a)k(a)})$ distribution. Setting $X'_a = Z_{k(a)} + E_a$ we observe that $\mathbf{X}'$ follows a $\mathcal{N}(0, \Sigma)$ Gaussian distribution, so $\mathbf{X} = \mathbf{X}'$ in distribution. □

1.1.4. *Proof of Proposition 5 of Section 2.5.2.* (i) We only need to prove $\Leftarrow$. If $\mathbf{X}$ has a G-block covariance, then for any $\sigma \in \mathcal{S}_G$ we have $cov(X_\sigma) = cov(X)$ so the distribution of X_σ is $\mathcal{N}(0, \Sigma)$. (ii) The second equivalence follows from Proposition 3 and Lemma 2. (iii) The first part follows immediately from above. For the second part, let $G^\ell(X)$ be a minimal element of $\mathcal{L}(X)$. Since $\mathcal{L}(X) \subseteq \mathcal{B}(X)$, then $G^\beta(X) \leq G^\ell(X)$. Since $\mathbf{X}$ is $G^\beta(X)$-latent, by Proposition 4, and $G^\ell(X)$ is minimal, we also have $G^\ell(X) \leq G^{block}(X)$. Therefore $G^\ell(X) = G^{block}(X)$ and the proof is complete. □

1.2. *Proof of Corollary 1 of Section 6.* Hoeffding concentration inequality for U-statistics [12] gives

$$\mathbf{P}\left(|\hat{\tau}_{ab} - \tau_{ab}| \geq t\right) \leq e^{-nt^2/8}, \quad \text{for all } a < b \text{ and } t \geq 0.$$

If $\widehat{R}$ is given by (5), and since $x \to \sin(\pi x/2)$ is $(\pi/2)$-Lipschitz, we obtain for all $t \geq 0$

$$\mathbf{P}\left(|\widehat{R} - R|_\infty \geq t\right) \leq p^2 e^{-nt^2/(2\pi^2)}.$$

Taking $t = 2\pi\sqrt{(1+A)\log(p)/n}$, the result then follows from Proposition 6, since $\tau \leq 2\pi\sqrt{(1+A)\frac{\log(p)}{n}}$ with probability higher than $1 - p^{-2A}$. □

1.3. *A corollary for Gaussian distributions .*

Corollary 1. *Let us consider $A > 0$ and (α, η) such that*

$$\alpha \geq 16\sqrt{\frac{(1+A)\log(p+1)}{n}} \quad \textit{and} \quad \eta \geq \alpha + 16\sqrt{\frac{(1+A)\log(p+1)}{n}}.$$

If $\mathbf{X}$ *has a zero mean Gaussian distribution with correlation matrix* $R \in \mathcal{R}(\eta)$*, then the output of our algorithm applied to the sample correlation matrix* $\widehat{R}$ *is consistent:* $\widehat{G} = G^{block}$*, with probability higher than* $1 - p^{-2A}$*.*

We remind the reader that the scaled variable $\mathbf{Y} = scale(\mathbf{X})$ is defined by $Y_a = X_a / (\Sigma_{aa})^{1/2}$, for $a = 1, \dots, p$. We write S for the covariance matrix of $\mathbf{Y}$, which coincides with the correlation matrix of $\mathbf{X}$, i.e. $S = R$. We also write $\widehat{S}$ for the (unobserved) empirical covariance matrix of $\mathbf{Y}$. The following lemma connects the sup-norm of $\widehat{R} - R$ to the sup-norm of $\widehat{S} - S$.

LEMMA 1. *We have* $|\widehat{R} - R|_\infty \leq 2|\widehat{S} - S|_\infty$.

Proof of the lemma. Since the empirical correlation matrices of $\mathbf{X}$ and $\mathbf{Y}$ are equal, we have $\widehat{R}_{ab} = \widehat{S}_{ab}/(\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}$. Since $R = S$, the triangle inequality gives

$$\begin{aligned} |\widehat{R}_{ab} - R_{ab}| &= |\widehat{R}_{ab}\big(1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}\big) + \widehat{S}_{ab} - S_{ab}| \\ &\leq |\widehat{R}_{ab}|\,|1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}| + |\widehat{S}_{ab} - S_{ab}|. \end{aligned}$$

We notice that

$$|1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}| \leq |1 - \widehat{S}_{aa}| \vee |1 - \widehat{S}_{bb}| = |S_{aa} - \widehat{S}_{aa}| \vee |S_{bb} - \widehat{S}_{bb}|.$$

Since $|\widehat{R}_{ab}| \leq 1$, we conclude that for any $a, b \in \{1, \dots, p\}$

$$\begin{aligned} |\widehat{R}_{ab} - R_{ab}| &\leq |\widehat{R}_{ab}| \left(|S_{aa} - \widehat{S}_{aa}| \vee |S_{bb} - \widehat{S}_{bb}|\right) + |\widehat{S}_{ab} - S_{ab}| \\ &\leq 2|\widehat{S} - S|_\infty. \end{aligned}$$

The proof of Lemma 1 is complete. □

When $16(1 + A)\log(p + 1) \leq n$, the classical bound on Gaussian covariance matrices (see e.g. [11], p.159)

$$\mathbf{P}(|\widehat{S} - S|_\infty > t) \leq p(p + 1)e^{-nt^2/8}, \text{ for any } 0 < t \leq 1,$$

ensures that $|\widehat{S} - S|_\infty \leq 4\sqrt{(1 + A)\log(p + 1)/n}$ with probability at least $1 - p^{-2A}$. Hence, according to Lemma 1, we have

$$|\widehat{R} - R|_\infty \leq 8\sqrt{(1 + A)\log(p + 1)/n} \tag{1}$$

with probability at least $1 - p^{-2A}$. Since $|\widehat{R} - R| \leq 2$, the bound (1) remains valid when $16(1 + A)\log(p + 1) > n$. The conclusion of Corollary 3 then follows from Proposition 6. □

2. Supplemental numerical results for simulated Gaussian data. We collect here the numerical results obtained in the Gaussian case. They are very close to those obtained for simulated Gaussian copula data. Figure 1 shows the percentage of exact recovery, the ARI index, and the performances of CV. Figure 2 displays the average number of False Discovery.

3. Supplemental results for the Standard & Poor 100 Stocks example. To illustrate the intuition behind the CORD metric, we compare the estimated correlation matrices corresponding to stocks, after reordering the variables representing the stocks by the group labels recovered by the three methods under evaluation. Because the group labels are arbitrary, to make the comparison as fair as possible, we use an approximate matching method [15] to match the labels of the other methods to each one of CORD. Figure 3 plots the estimated Kendall's tau correlation matrices after re-ordering by the matched labels. Visually, it is clear to see that the CORD partition shows mosaic block patterns for both the diagonal and off-diagonal parts of the correlation matrix. The other methods with criteria based on intra-class distances mainly show diagonal blocks, and there could be multiple bands within each group because the variables in groups obtained by either K-means or HC can have different correlations with all other variables, whereas for CORD, variables in a group have the same, high or low, correlation with all other variables.

4. A functional MRI example. Using functional MRI data, [23] found that the human brain regions are organized into communities, sometimes referred to as networks. We use a publicly available fMRI dataset to illustrate the communities recovered by CORD. The dataset was originally published in [26] and is available from Open fMRI (https://openfmri.org/data-sets) under the accession number ds000007. We will focus on analyzing two scan sessions from subject 1 under a visual-motor stop/go task (task 1). Before performing the analysis, we follow the preprocessing steps suggested by [26], see Section 4.1 below for details.

Using the first run data alone, the same CV procedure yields $K = 80$ with $\alpha = 1.5n^{-1/2}\log^{1/2} p$, close to the fixed constant 2 in our simulations. We thus set $K = 80$ in K-means and HC, and use the same procedure to match the groups across the methods. The correlation matrices after reordering the variables are plotted in Figure 4. By visual comparison, the CORD groups are aligned with the bands in both rows and columns of the correlation matrix, while the other methods focusing on recovering diagonal blocks can have multiple bands in the same group.

Figure 1: **Top:** Percentages of exact recovery by K-means (medium blue lines, triangle points), HC (light yellow lines, cross points) and CORD (dark red lines, circle points) across 100 runs, when $p = 200$ (solid lines) and $p = 1600$ (dashed lines). The minimal sample sizes are shown by light red vertical lines. All standard errors are smaller than 5%. **Center:** Average ARI by K-means (medium blue lines, triangle points), HC (light yellow lines, cross points) and CORD (dark red lines, circle points) across 100 runs. **Bottom:** Average CV losses (black dotted lines, plus points) and exact recovery percentages (red solid lines, circle points) across 100 runs. The minimal average CV losses are shown by gray vertical lines.

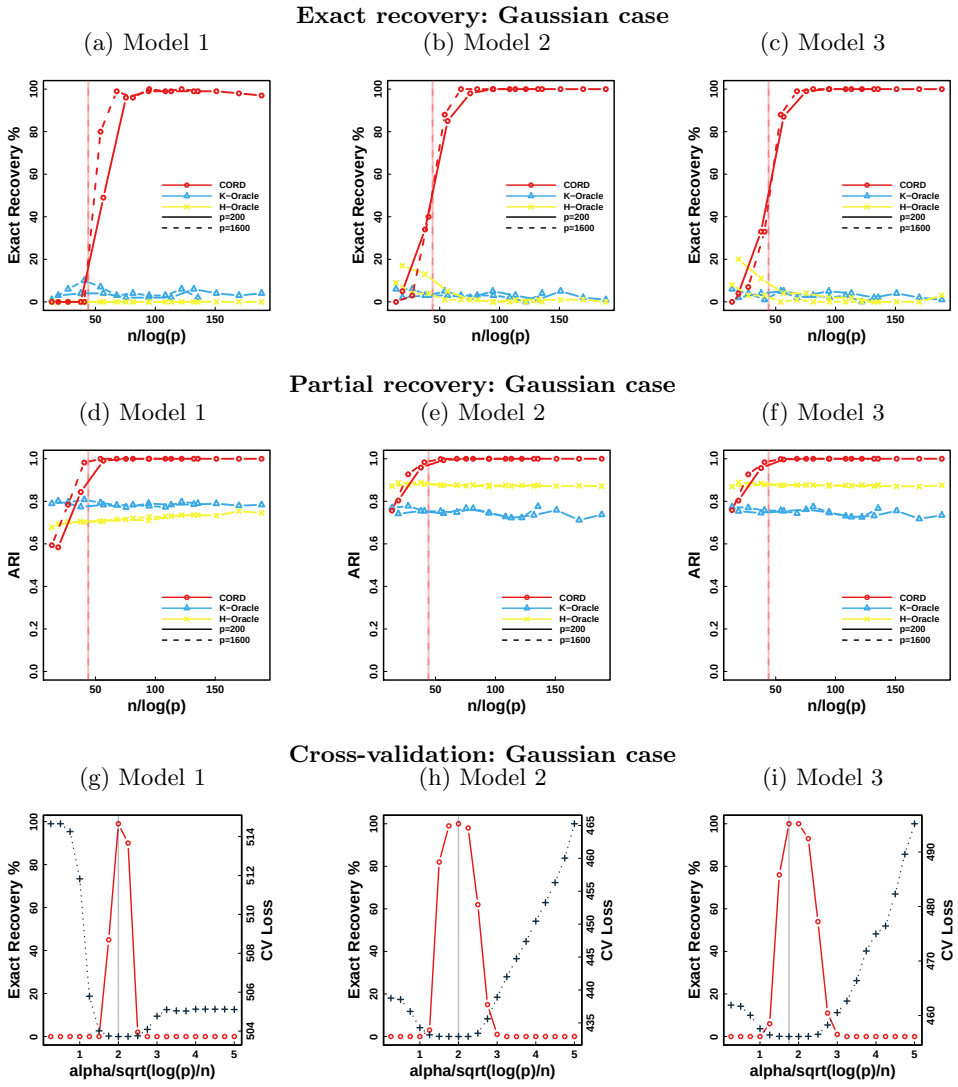


Figure 2: Average numbers of false discoveries (FD) and discoveries (D) after thresholding either the CORD (red solid lines) or correlation (blue dotted lines) metrics. The shaded areas show mean±SD.

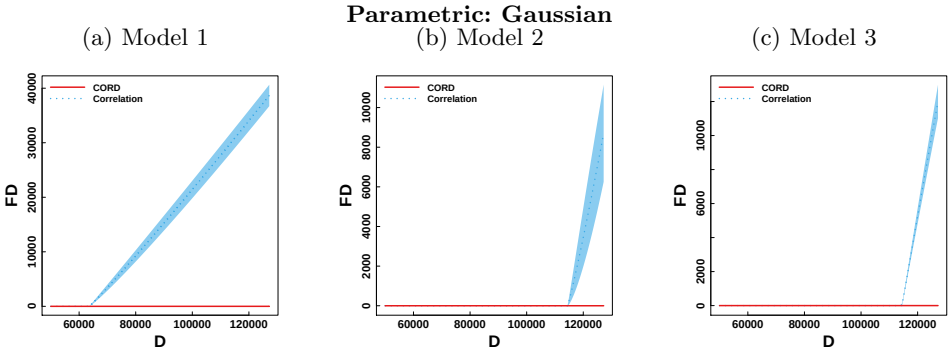


Figure 3: Comparison of correlation matrices (black=1) when the variables are ordered by the groups recovered by (a) CORD, (b) Kmeans, and (c) HC. The variable groups are denoted by color sidebars.

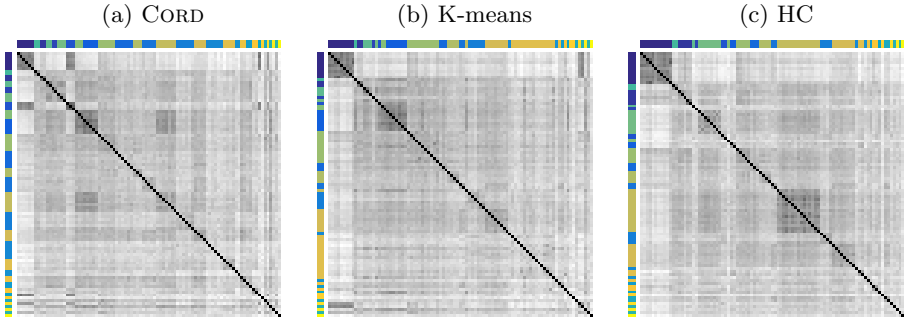
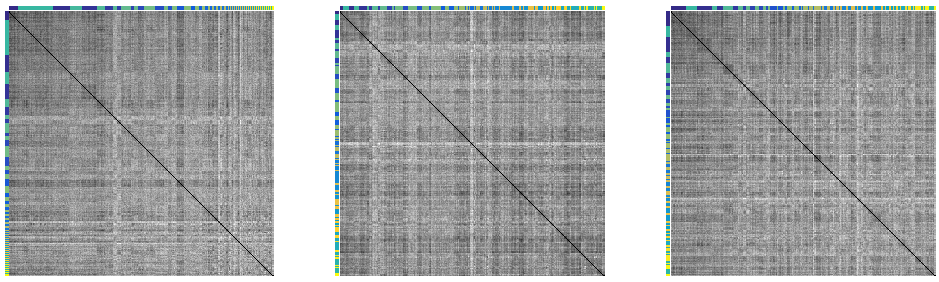


Figure 4: Heatmaps (black=1) of correlation matrices after re-ordering the variables according to the recovered groups by CORD, K-means, and HC. The order of groups from K-means and HC are approximately matched to that of CORD, and each group is denoted by color using the side bars.

(a) CORD (b) K-means (c) HC



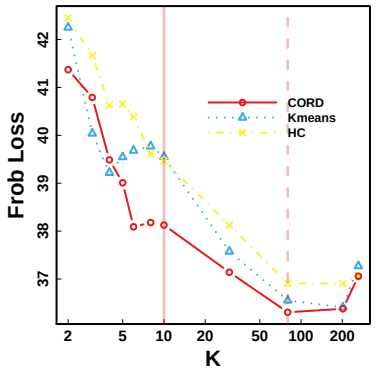
The largest CORD group contain 35 ROIs. We use the brain function classification map available at http://www.brainnexus.com/resources/resting-state-fmri-templates to find the functioning of these ROIs. This CORD group includes 14 visual areas, 14 motor areas, 4 salience network areas, and 3 executive areas. The majority of the ROIs in this cluster correspond to visual-motor functioning, which is well expected to coordinate together during the visual-motor task. The salience and executive areas are also expected to be involved during the task, see a review [25]. We refer to Section 4.1 below for additional results.

Because there are no gold standards for partitioning the brain, we use a prediction criterion to compare CORD with K-means and HC. We compute the correlation matrices $\hat{R}_1$ and $\hat{R}_2$ from the first and second session data respectively. For a grouping estimate $\hat{G}$, We use the following loss to evaluate its performance

$$\left\| \hat{R}_2 - \Upsilon\left(\hat{R}_1, \hat{G}\right) \right\|_F. \tag{2}$$

For a fair comparison, we show in Figure 5 the losses against different group sizes, where the groups are estimated by CORD, Kmeans, and HC respectively. Here, we use Kendal's τ for computing the correlation matrices, and the results for Pearson's correlations are similar. Again, CORD outperforms all other methods for almost all group sizes. Compared with K-means and HC, CORD yields the smallest loss values for a wide range of K, and the CV selected $K = 80$ yields the smallest loss.

Figure 5: Comparison of CORD, K-means, and HC using two session data, using the Frobenius prediction loss criterion (2) where the groups are estimated by these methods respectively. The group sizes for the fixed and CV choices are shown as solid and dashed vertical lines respectively.



4.1. *Materials on the fMRI data analysis.*

Preprocessing. We applied the preprocessing steps suggested by [26], which includes slice timing correction, alignment, registration, normalization to the average 152 T1 MNI template, smoothing with a 5mm full-width-half-maximum Gaussian kernel, denoising using the FSL MELODIC procedure, and a high pass filter with a 66s cut-off. The event-related activation and temporal correlation were removed using general linear models (GLM) for each voxel [9]. Following [23], we extract 180 mean activities within a 10mm spheres centered around each of 264 putative functional areas (see Table S2 from [23]).

Additional results. We plot in Figure 6 the brain ROIs and their CORD clustering memberships.
The MNI coordinates (in mm) of the ROIs in the largest CORD are included in Table 1.

REFERENCES

[1] Arias-Castro, E. and Verzelen, N. (2013) Community detection in random networks. Preprint.

[2] Amini, A.A., Chen, A., Bickel, P.J., and Levina, E. (2013) Pseudo likelihood metgods for community detection in large sparse networks. Annals of Statistics 41 (4), 2097 - 2122.

[3] Bunea, F., Giraud, C., and Luo, X., 2015. Minimax optimal variable clustering in G-models via CORD. Preprint.

Figure 6: Brain ROIs and their CORD groups (shown by color nodes) under the (a) sagittal, (b) axial, and (c) coronal views.

(a) Sagittal (b) Axial (c) Coronal

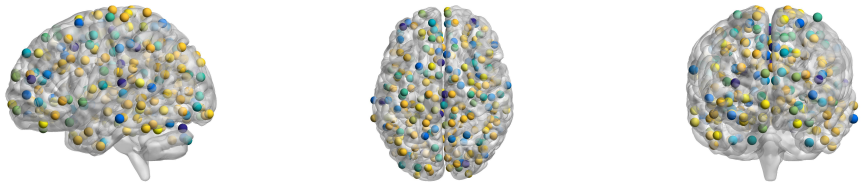


TABLE 1
MNI coordinates (x, y, z, in mm) of the largest CORD *group and their functioning classification.*

X	Y	Z	Function	X	Y	Z	Function
6	-59	35	visual	38	-17	45	motor
11	-54	17	visual	-49	-11	35	motor
-12	-95	-13	visual	51	-6	32	motor
8	-72	11	visual	-5	18	34	motor
-8	-81	7	visual	58	-16	7	motor
27	-59	-9	visual	-38	-33	17	motor
-18	-68	5	visual	52	-59	36	motor
-47	-76	-10	visual	-28	-58	48	motor
15	-77	31	visual	-10	-18	7	motor
-16	-52	-1	visual	12	-17	8	motor
6	-72	24	visual	31	-14	2	motor
-42	-74	0	visual	29	1	4	motor
-16	-77	34	visual	9	-4	6	motor
6	-81	6	visual	22	-58	-23	motor
-11	45	8	salience	-58	-30	-4	executive
-53	-49	43	salience	35	-67	-34	executive
0	30	27	salience	-47	-51	-21	executive
26	50	27	salience				

[4] Bunea, F. and Xiao, L, (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. Bernoulli 21(2), 1200-1230.

[5] Chen, Y. and Xu, J. (2015) Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices, JMLR, to appear.

[6] Christensen, D. (2005). Fast algorithms for the calculation of Kendall's τ. Computational Statistics 20 (1): 51–62.

[7] Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human brain mapping, 33(8), 1914-1928.

[8] Craddock, R. C., Jbabdi, S., Yan, C. G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., ... and Milham, M. P. (2013). Imaging human connectomes at the macroscale. Nature methods, 10(6), 524-539.

[9] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping, 2(4), 189-210.

[10] Gao,C., Ma, Z., Zhang, A. y. and Zhou, H. (2015) Achieving optimal misclassification proportion in stochastic block model, Preprint.

[11] Giraud, C. (2014) Introduction to High-Dimensional Statistics. Chapman and Hall.

[12] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. JASA **58**: 13–30.

[13] Hartigan, J.A. (1975) Clustering algorithms. John Wiley & Sons.

[14] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983) Stochastic blockmodels: First steps. Social Networks 5, 109- 137.

[15] Hornik, K. (2005). A clue for cluster ensembles. Journal of Statistical Software, 14(12).

[16] Hubert, L., and Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.

[17] Izenman, A. (2008) Modern Multivariate Statistical Techniques. Springer Text in Statistics.

[18] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[19] Klopp, O., Tsybakov, A. B. and Verzelen, N. (2015) Oracle inequalities for network models and sparse graphon estimation.

[20] Lei, J. and Rinaldo, A. (2015) Consistency of spectral clustering in stochastic block models. Annals of Statistics 43(1), 215-237.

[21] Massart, P. (2007) Concentration Inequalities and Model Selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics 1896, Springer.

[22] Murtagh, F., and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.

[23] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., ... and Petersen, S. E. (2011). Functional network organization of the human brain. Neuron, 72(4), 665-678.

[24] Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006). Learning the structure of linear latent variable models. The Journal of Machine Learning Research, 7, 191-246.

[25] Simmonds, D. J., Pekar, J. J., and Mostofsky, S. H. (2008). Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. Neuropsychologia, 46(1), 224-232.

[26] Xue, G.; Aron, A. R. and Poldrack, R. A.(2008). Common neural substrates for inhibition of spoken and manual responses Cerebral Cortex, Oxford Univ Press,, 18, 1923-1932.

Department of Statistical Science
Cornell University
Ithaca, NY 14853-2601, USA
E-mail: fb238@cornell.edu

Département de Mathématiques
Université Paris-Sud
F-91405 Orsay Cedex, France
E-mail: christophe.giraud@math.u-psud.fr

Department of Biostatistics
and Center for Statistical Science
Brown University
Providence, RI 02912, USA
E-mail: xi.rossi.luo@gmail.com